

Markov Chain,Hidden Markov Model

April 5, 2017

Discrete-time Markov chains

- Definition: A Markov chain is a discrete-time stochastic process $(X_n, n \geq 0)$ such that each random variable X_n takes values in a discrete set S ($S = N$, typically) and

$$P(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = P(X_{n+1} = j | X_n = i),$$

$$\forall n \geq 0, j, i, i_{n-1}, \dots, i_0 \in S$$

- If $P(X_{n+1} = j | X_n = i) = p_{ij}$ is independent of n , then X is said to be a time-homogeneous Markov chain.

- Terminology

The possible values taken by the random variables X_n are called the states of the chain. S is called the state space.

The chain is said to be finite-state if the set S is finite ($S = \{0, \dots, N\}$, typically).

$P = (p_{ij})$ $i, j \in S$ is called the transition matrix of the chain.

- Properties of the transition matrix

$$p_{ij} \geq 0, \forall i, j \in S$$

$$\sum_{\forall j \in S} p_{ij} = 1, \forall i \in S.$$

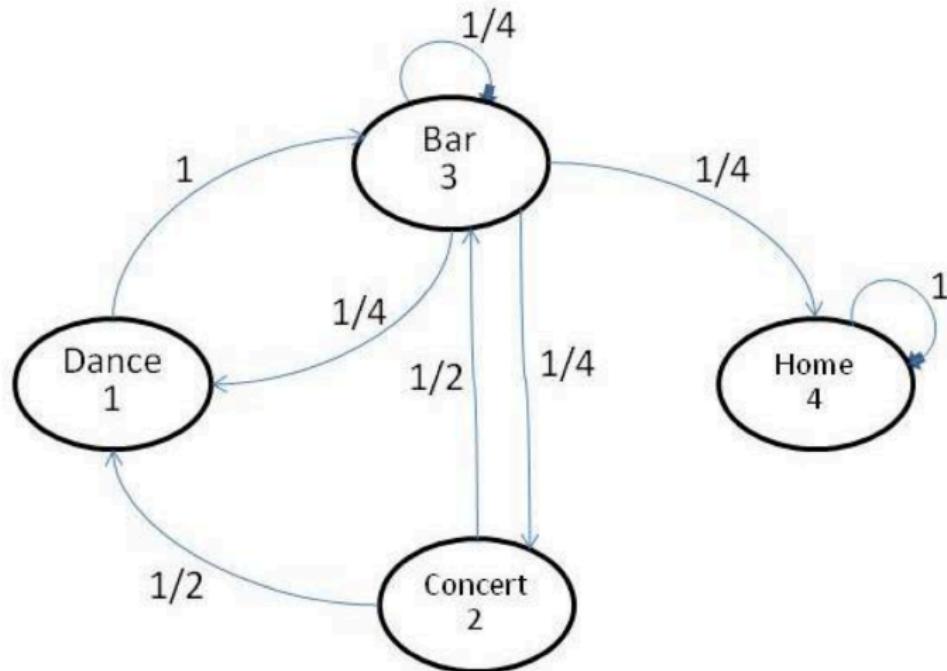
Example 1: Music Festival

- The four possible states of a student in a music festival are $S = \text{"dancing", "at a concert", "at the bar", "back home"}$. Let us assume that the student changes state during the festival according to the following transition matrix:

$$P = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Example

The Markov chain can be represented by the following transition graph:



Example 2: Simple symmetric random walk

- Let $(X_n, n \geq 1)$ be i.i.d. random variables such that $P(X_n = +1) = P(X_n = -1) = 1/2$, and let $(S_n, n \geq 0)$ be defined as $S_0 = 0, S_n = X_1 + \dots + X_n, n \geq 1$. Then $(S_n, n \in N)$ is a Markov chain with state space $S = Z$.

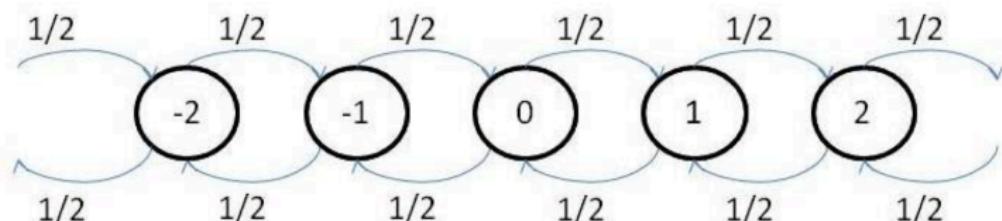
- Indeed:

$$P(S_{n+1} = j | S_n = i, S_{n-1} = i_{n-1}, \dots, S_0 = i_0) \\ = P(X_{n+1} = j - i | S_n = i, S_{n-1} = i_{n-1}, \dots, S_0 = i_0) = P(X_{n+1} = j - i)$$

by the assumption that the variables X_n are independent.

- The chain is time-homogeneous, as $P(X_{n+1} = j - i) = 1/2$, if $|j - i| = 1$, $P(X_{n+1} = j - i) = 0$, otherwise, which does not depend on n .

The transition graph of the chain:



- The distribution at time n of the Markov chain X is given by:

$$\pi_i^{(n)} = P(X_n = i), i \in S.$$

- We know that $\pi_i^{(n)} \geq 0 \forall i \in S$ and that $\sum_{i \in S} \pi_i^{(n)} = 1$.
- The initial distribution of the chain is given by
$$\pi_i^{(0)} = P(X_0 = i), i \in S.$$
- It must be specified together with the transition matrix
 $P = (p_{ij}), i, j \in S$ in order to characterize the chain completely.

$$\begin{aligned}
 & P(X_n = i_n, X_{n-1} = i_{n-1}, \dots, X_1 = i_1, X_0 = i_0) \\
 = & P(X_n = i_n | X_{n-1} = i_{n-1}, \dots, X_1 = i_1, X_0 = i_0) \cdot \\
 & P(X_{n-1} = i_{n-1}, \dots, X_1 = i_1, X_0 = i_0) \\
 = & p_{i_{n-1}, i_n} P(X_{n-1} = i_{n-1}, \dots, X_1 = i_1, X_0 = i_0) \\
 = & \dots = p_{i_{n-1}, i_n} p_{i_{n-2}, i_{n-1}} \cdots p_{i_1, i_2} p_{i_0, i_1} \pi(0)
 \end{aligned}$$

So knowing P and knowing $\pi(0)$ allows to compute all the above probabilities, which give a complete description of the process.

The n -step transition probabilities of the chain are given by

$$p_{ij}^{(n)} = P(X_{m+n} = j | X_m = i), n, m \geq 0, i, j \in S.$$

Let us compute:

$$\begin{aligned} p_{ij}^{(2)} &= P(X_{n+2} = j | X_n = i) = \sum_{k \in S} P(X_{n+2} = j, X_{n+1} = k | X_n = i) \\ &= \sum_{k \in S} P(X_{n+2} = j | X_{n+1} = k, X_n = i) \cdot P(X_{n+1} = k | X_n = i) \\ &= \sum_{k \in S} P(X_{n+2} = j | X_{n+1} = k) \cdot P(X_{n+1} = k | X_n = i) \\ &= \sum_{k \in S} p_{ik} p_{kj} \end{aligned}$$

The Chapman-Kolmogorov equation for generic values of m and n :

$$p_{ij}^{(n+m)} = \sum_{k \in S} p_{ik}^{(n)} p_{kj}^{(m)}, i, j \in S, n, m \geq 0,$$

where we define by convention $p_{ij}^{(0)} = \delta_{ij} = 1$, if $i = j$, $p_{ij}^{(0)} = \delta_{ij} = 0$ otherwise.

In matrix form:

$$P_{ij}^{n+m} = (P^n P^m)_{ij} = \sum_{k \in S} (P^n)_{ik} (P^m)_{kj}$$

$$\pi^{(n)} = \pi^{(n-1)} P$$

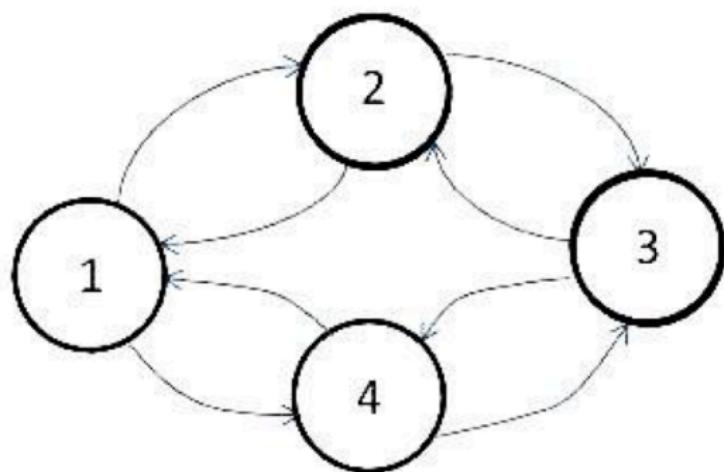
$$\pi_j^{(n)} = \sum_{i \in S} p_{ij}^{(n)} \pi_i^{(0)}$$

Classification of states

- A state j is accessible from state i if $p_{ij}^{(n)} > 0$ for some $n \geq 0$.
- State i and j communicate if both j is accessible from i and i is accessible from j .
- Two states that communicate are said to belong to the same equivalence class, and the state space S is divided into a certain number of such classes.
- The Markov chain is said to be irreducible if there is only one equivalence class (i.e. all states communicate with each other).
- A state i is absorbing if $p_{ii} = 1$.
- A state i is periodic with period d if d is the smallest integer such that $p_{ii}^{(n)} = 0$ for all n which are not multiples of d .
- In case $d = 1$, the state is said to be aperiodic.

Example

The Markov chain whose transition graph is given by is an irreducible Markov chain, periodic with period 2.



Recurrent and transient states

- Let us also define $f_i = P(X \text{ ever returns to } i | X_0 = i)$. A state i is said to be recurrent if $f_i = 1$ or transient iff $f_i < 1$.
- It can be shown that all states in a given class are either recurrent or transient.
- In Example musical concert, the class “dancing”, “at a concert”, “at the bar” is transient (as there is a positive probability to leave the class and never come back) and the class “back home” is obviously recurrent.

$$P = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

- Definition:

Let i be a recurrent state and T_i be the first return time to state i .

i is positive recurrent if $E(T_i|X_0 = i) < \infty$

i is null recurrent if $E(T_i|X_0 = i) = \infty$.

- Proposition:

State i is recurrent if and only if $\sum_{n \geq 1} p_{ii}^{(n)} = \infty$.

State i is transient if and only if $\sum_{n \geq 1} p_{ii}^{(n)} < \infty$.

- Proof. Let T_i be the first time the chain X returns to state i
 $f_i = P(T_i < \infty | X_0 = i)$. Let N_i be the number of times that the chain returns to state i . Then

$$\begin{aligned} P(N_i < \infty | X_0 = i) &= \sum_{n \geq 1} P(X_n = i, X_m \neq i, \forall m > n | X_0 = i) \\ &= \sum_{n \geq 1} P(X_m \neq i, \forall m > n | X_n = i, X_0 = i) P(X_n = i | X_0 = i) \end{aligned}$$

As X is a time-homogeneous Markov chain, we have

$$\begin{aligned} P(X_m \neq i, \forall m > n | X_n = i, X_0 = i) &= P(X_m \neq i, \forall m > n | X_n = i) \\ &= P(X_m \neq i, \forall m > 0 | X_0 = i) \end{aligned}$$

Therefore,

$$\begin{aligned} P(N_i < \infty | X_0 = i) &= P(X_m \neq i, \forall m > 0 | X_0 = i) \sum_{n \geq 1} P(X_n = i | X_0 = i) \\ &= P(T_i = \infty | X_0 = i) \sum_{n \geq 1} p_{ii}^{(n)} = (1 - f_i) \sum_{n \geq 1} p_{ii}^{(n)} \end{aligned}$$

- If i is recurrent, then $1 - f_i = 0$. We have $P(N_i < \infty | X_0 = i) = 0$. This in turn implies

$$P(N_i = \infty | X_0 = i) = 1, \text{ so } E(N_i | X_0 = i) = \infty.$$

As $N_i = \sum_{n \geq 1} \mathbf{1}_{\{X_n=i\}}$, we have

$$\begin{aligned}\infty &= E(N_i | X_0 = i) = \sum_{n \geq 1} E(\mathbf{1}_{\{X_n=i\}} | X_0 = i) \\ &= \sum_{n \geq 1} P(X_n = i | X_0 = i) = \sum_{n \geq 1} p_{ii}^{(n)}\end{aligned}$$

- If i is transient, then $1 - f_i > 0$. As $P(N_i = \infty | X_0 = i) \leq 1$, we obtain

$$\sum_{n \geq 1} p_{ii}^{(n)}(1 - f_i) \leq 1, \sum_{n \geq 1} p_{ii}^{(n)} < \infty.$$

Stationary and limiting distributions

- A distribution $\pi^* = (\pi_i^*, i \in S)$ is said to be a stationary distribution for the Markov chain $(X_n, n \geq 0)$ if

$$\pi^* = \pi^* P, \text{ i.e. } \pi_j^* = \sum_{i \in S} \pi_i^* \cdot p_{ij}, \forall j \in S$$

- π^* does not necessarily exist, nor is it necessarily unique.
- If π^* exists and is unique, then π_i^* can always be interpreted as the average proportion of time spent by the chain X in state i .
 $E(T_i | X_0 = i) = 1/\pi_i^*$, where $T_i = \inf\{n \geq 0 : X_n = i\}$ is the first time the chain comes back to state i .
- If $\pi^{(0)} = \pi^*$, then $\pi^{(1)} = \pi^* P = \pi^*$; likewise, $\pi^{(n)} = \pi^*, \forall n \geq 0$.

- A distribution $\pi^* = (\pi_i^*, i \in S)$ is said to be a limiting distribution for the Markov chain $(X_n, n \geq 0)$ if for every initial distribution $\pi^{(0)}$ of the chain, we have

$$\lim_{n \rightarrow \infty} \pi^{(n)} = \pi^*, \forall i \in S$$

- If π^* is a limiting distribution, then it is stationary. Indeed, for all $n \geq 0$, we always have $\pi^{(n+1)} = \pi^{(n)}P$. If $\lim_{n \rightarrow \infty} \pi^{(n)} = \pi^*$, then $\pi^* = \pi^*P$.
- A limiting distribution π^* does not necessarily exist, but if it exists, then it is unique.

- Theorem 1. Let $(X_n, n \geq 0)$ be an irreducible and aperiodic Markov chain. Let us assume that it admits a stationary distribution π^* . Then π^* is a limiting distribution, i.e. for any initial distribution $\pi^{(0)}$, $\lim_{n \rightarrow \infty} \pi_i^{(n)} = \pi_i^*, \forall i \in S$.
- Theorem 2. Let $(X_n, n \geq 0)$ be an irreducible and positive recurrent Markov chain. Then X admits a unique stationary distribution π^* .

- Definition. A (time-homogeneous) Markov chain $(X_n, n \geq 0)$ is said to be ergodic if it is irreducible, aperiodic and positive recurrent.
- Corollary. An ergodic Markov chain $(X_n, n \geq 0)$ admits a unique stationary distribution π^* . Moreover, this distribution is also a limiting distribution, i.e. $\lim_{n \rightarrow \infty} \pi^{(n)} = \pi^*, \forall i \in S$.

Proof

证：有限 Markov 链是遍历的，故存在一个正整数 m ，使得对于状态空间中的任何状态 i, j 有 $p_{i,j}^{(m)} > 0$ 。下面证明对于转移 M ，有 $\lim_{n \rightarrow \infty} M^{(n)} = \pi$ 。这里 π 是一个随机矩阵，且它的各行都相同，它的每行即是平稳分布。

1、先证明 $m=1$ 的情形。

若 $m=1$ ，则由条件可得， $p_{i,j} \geq \varepsilon > 0$ 。令 $m_j(n) = \min_i p_{i,j}^{(n)}$ ， $M_j(n) = \max_i p_{i,j}^{(n)}$ 。

由 Chapman-Kolmogorov 方程可知， $p_{i,j}^{(n)} = \sum_k p_{i,k} p_{k,j}^{(n-1)} \geq \sum_k p_{i,k} m_j(n-1) = m_j(n-1)$ 。

上式对于所有的 i 都成立，因此， $m_j(n) \geq m_j(n-1)$ ，这说明 $m_j(n)$ 随着 n 的增加而增加。

同理可以说明 $M_j(n)$ 随着 n 的增加而减少，即 $M_j(n) \leq M_j(n-1)$ 。由于， $m_j(n)$ 、 $M_j(n)$ 都是有界序列，故它们的极限存在。下面只需证明这两个序列的极限相同。

设当 $i = i_0$ 时，经 n 步转移后到达最小值 $m_j(n)$ ，又设 $i = i_1$ 时，经 $n-1$ 步转移后到达最

Proof

$$\begin{aligned} \text{大值 } M_j(n-1), \text{ 则 } m_j(n) &= p_{i_0,j}^{(n)} = \sum_k p_{i_0,k} p_{k,j}^{(n-1)} \\ &= \varepsilon p_{i_1,j}^{(n-1)} + (p_{i_0,i_1} - \varepsilon) p_{i_1,j}^{(n-1)} + \sum_{k \neq j} p_{i_0,k} p_{k,j}^{(n-1)} \\ &\geq \varepsilon M_j(n-1) + [p_{i_0,i_1} - \varepsilon + \sum_{k \neq j} p_{i_0,k}] m_j(n-1) \end{aligned}$$

即 $m_j(n) \geq \varepsilon M_j(n-1) + (1-\varepsilon) m_j(n-1)$ 。

同理可得， $M_j(n) \leq \varepsilon m_j(n-1) + (1-\varepsilon) M_j(n-1)$ 。

上式两式相减， 得 $M_j(n) - m_j(n) \leq (1-2\varepsilon)[M_j(n-1) - m_j(n-1)]$ ， 递推可得，

$M_j(n) - m_j(n) \leq (1-2\varepsilon)^{n-1}$ 。因此， $n \rightarrow \infty$ 时， $m_j(n)$ 、 $M_j(n)$ 趋于同一极限。这就证明

了 $\lim_{n \rightarrow \infty} M^{(n)} = \pi$ 。 π 是各行都相同的矩阵。

Proof

2、 $m > 1$ 时， $\lim_{n \rightarrow \infty} (M^{(m)})^{(n)} = \lim_{n \rightarrow \infty} M^{(mn)} = \pi$ 。又设 $k=1, 2, 3, \dots, m-1$ ，则

$\lim_{n \rightarrow \infty} M^{(mn+k)} = \lim_{n \rightarrow \infty} M^{(mn)} M^{(k)} = M^{(k)} \pi$ 。因为在 $M^{(k)}$ 中各行之和为 1，且 π 中的任何一列的元素都相等，因此 $M^{(k)} \pi = \pi$ ，即 $\lim_{n \rightarrow \infty} M^{(mn+k)} = \pi$ 。

3、下面说明 $\lim_{n \rightarrow \infty} P(\xi_n = j) = \lim_{n \rightarrow \infty} p_{i,j}^{(n)} = \pi_j$ ，即 $\lim_{n \rightarrow \infty} P(\xi_n = j)$ 所取的值与初始状态的分布无关。 $P(\xi_n = j) = \sum_i P(\xi_n = j | \xi_0 = i) P(\xi_0 = i) = \sum_i p_{i,j}^{(n)} P(\xi_0 = i)$ ，

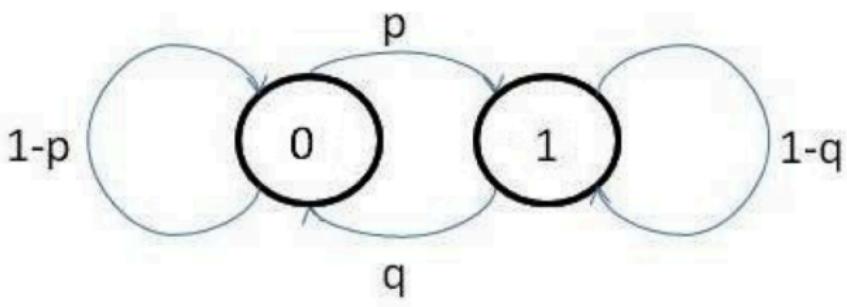
$$\begin{aligned} \text{故 } \lim_{n \rightarrow \infty} P(\xi_n = j) &= \lim_{n \rightarrow \infty} \sum_i p_{i,j}^{(n)} P(\xi_0 = i) = \sum_i \pi_j P(\xi_0 = i) \\ &= \pi_j \sum_i P(\xi_0 = i) = \lim_{n \rightarrow \infty} p_{i,j}^{(n)} \end{aligned}$$

4、证明其唯一性。

如果另有一矩阵 W 也满足 $WM=W$ ，且 W 的各行之和为 1，则

$W = WM = WM^{(2)} = \dots = WM^{(n)}$ 。当 $n \rightarrow \infty$ 时， $W = W\pi$ 。 W 中每行之和为 1，且 π 中每列元素都相等，从而 $W = W\pi = \pi$ 。唯一性得证。

Example: two-state Markov chain

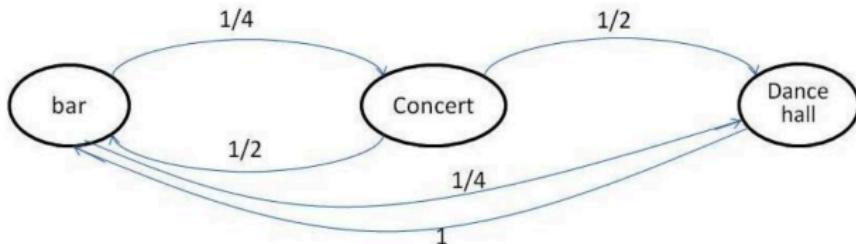


- Both $p, q > 0$, this chain is clearly irreducible, and as it is finite-state, it is also positive recurrent. It admits a stationary distribution.
- Since $\pi = \pi P$, we obtain

$$\pi_0 = \pi_0(1 - p) + \pi_1 q, \pi_1 = \pi_0 p + \pi_1(1 - q)$$

- π is a distribution, we must have $\pi_0 + \pi_1 = 1$ and $\pi_0, \pi_1 \geq 0$.
- Solving these equations, we obtain $\pi_0 = q/(p + q)$, $\pi_1 = p/(p + q)$.
- If $p + q < 2$, then the chain is also aperiodic and therefore ergodic, so $\pi = (q/(p + q), p/(p + q))$ is also a limiting distribution.

Example:music festival



It has the corresponding transition matrix:

$$P = \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 1/2 & 0 & 1/2 \\ 1 & 0 & 0 \end{pmatrix}$$

We can again easily see that the chain is ergodic. The computation of its stationary and limiting distribution gives

$$\pi^* = (8/13, 2/13, 3/13)$$

Example:simple symmetric random walk

Let us consider the simple symmetric random walk. This chain is irreducible, periodic with period 2 and all states are null recurrent. There does not exist a stationary distribution here.

Proposition: Let $(X_n, n \geq 0)$ be a finite-state irreducible Markov chain with state space $S = \{0, \dots, N\}$ and let π^* be its unique stationary distribution. Then π^* is the uniform distribution if and only if the transition matrix P of the chain satisfies:

$$\sum_{i \in S} p_{ij} = 1, \forall j \in S.$$

Reversible Markov chains

Let $(X_n, n \geq 0)$ be a time-homogeneous Markov chain. Let us now consider this chain backwards, i.e. consider the process $(X_n, X_{n-1}, X_{n-2}, \dots, X_1, X_0)$: this process turns out to be also a Markov chain (but not necessarily time-homogeneous). Indeed:

$$\begin{aligned} & P(X_n = j | X_{n+1} = i, X_{n+2} = k, X_{n+3} = l, \dots) \\ = & \frac{P(X_n = j, X_{n+1} = i, X_{n+2} = k, X_{n+3} = l, \dots)}{P(X_{n+1} = i, X_{n+2} = k, X_{n+3} = l, \dots)} \\ = & \frac{P(X_{n+2} = k, X_{n+3} = l, \dots, | X_{n+1} = i, X_n = j)P(X_{n+1} = j, X_n = j)}{P(X_{n+2} = k, X_{n+3} = l, \dots, | X_{n+1} = i)P(X_{n+1} = i)} \\ = & \frac{P(X_{n+2} = k, X_{n+3} = l, \dots, | X_{n+1} = i)}{P(X_{n+2} = k, X_{n+3} = l, \dots, | X_{n+1} = i)} P(X_n = j | X_{n+1} = i) \\ = & P(X_n = j | X_{n+1} = i) \end{aligned}$$

Let us now compute the transition probabilities:

$$\begin{aligned} P(X_n = j | X_{n+1} = i) &= \frac{P(X_n = j, X_{n+1} = i)}{P(X_{n+1} = i)} \\ &= \frac{P(X_{n+1} = i | X_n = j)P(X_n = j)}{P(X_{n+1} = i)} \\ &= \frac{p_{ji}\pi_j^{(n)}}{\pi_i^{(n+1)}} \end{aligned}$$

The backward chain is not necessarily time-homogeneous.

- Let us now assume that the chain is irreducible and positive recurrent. Then it admits a unique stationary distribution π^* . Let us moreover assume that the initial distribution of the chain is the stationary distribution.

$$P(X_n = j | X_{n+1} = i) = \frac{p_{ji}\pi_j^*}{\pi_i^*} = \tilde{p}_{ij}$$

The backward chain is time-homogeneous with transition probabilities \tilde{p}_{ij} .

- Definition: The chain X is said to be reversible if $\tilde{p}_{ij} = p_{ij}$, i.e. the transition probabilities of the forward and the backward chains are equal. In this case, the following detailed balance equation is satisfied:

$$\pi_i^* p_{ij} = \pi_j^* p_{ji}, \forall i, j \in S.$$

- If a distribution π^* satisfies the above detailed balance equation, then it is a stationary distribution. Indeed,

$$\sum_{i \in S} \pi_i^* p_{ij} = \sum_{i \in S} \pi_j^* p_{ji} = \pi_j^* \sum_{i \in S} p_{ji} = \pi_j^*, \forall j \in S$$

- In order to find the stationary distribution of a chain, solving the detailed balance equation is easier than solving the stationary distribution equation, but this works of course only if the chain is reversible.
- The detailed balance equation has the following interpretation: it says that in the Markov chain, the flow from state i to state j is equal to that from state j to state i .
- If the detailed balance equation is satisfied, then π^* is the uniform distribution if and only if P is a symmetric matrix.

Parameter estimation for Markov chains

- The elements of the training set $\{x_1, \dots, x_n\}$, are assumed to be independent,

$$P(x_1, \dots, x_n | \theta) = \prod_j P(x_j | \theta).$$

- MLE parameter estimation looks for θ which maximizes the above.
- WLOG, we consider there is only one Markov chain X with length L . We have

$$P(X | \theta) = \prod_{j=1}^L p_{i_{j-1}, i_j}$$

Parameter estimation for Markov chains

- Let the transition probability matrix be P . $m_{kl} = |j : i_{j-1=k}, i_j = l|$. Then our optimization problem is formulated as:

$$\max P(X|\theta) = \prod_{k,l}^L p_{k,l}^{m_{kl}}$$

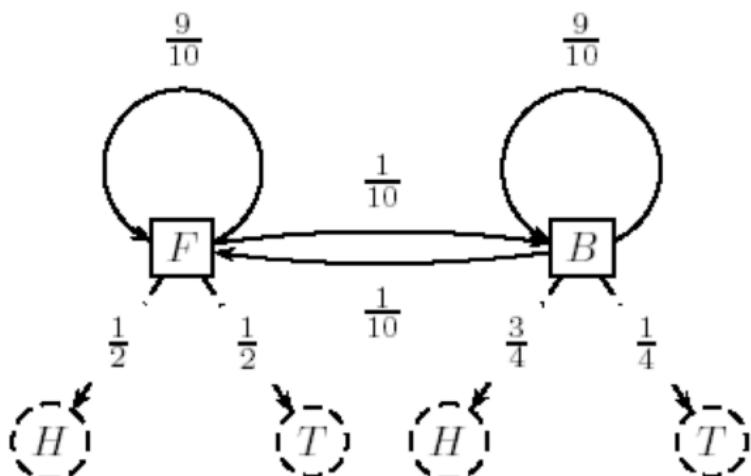
s.t.

$$\sum_l p_{k,l} = 1, \forall k.$$

- By solving the maximization problem, we can get the MLE is given by:

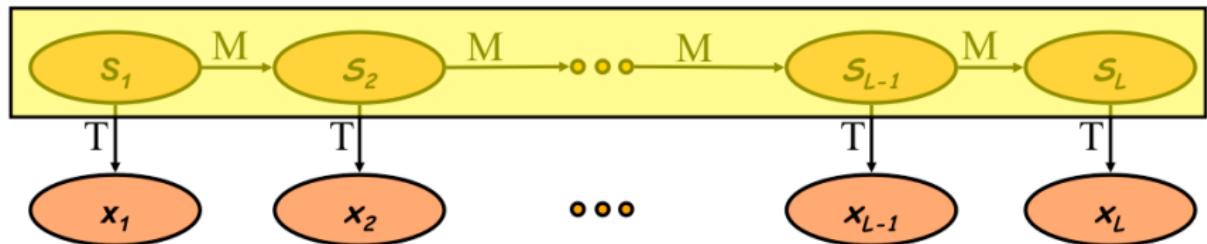
$$p_{k,l} = \frac{m_{kl}}{\sum_{l'} m_{kl'}}$$

Hidden Markov Model



HMM model for the *Fair Bet Casino Problem*

Hidden Markov Model



- Notations:

Markov Chain transition probabilities: $P(S_{i+1} = t | S_i = s) = a_{st}$

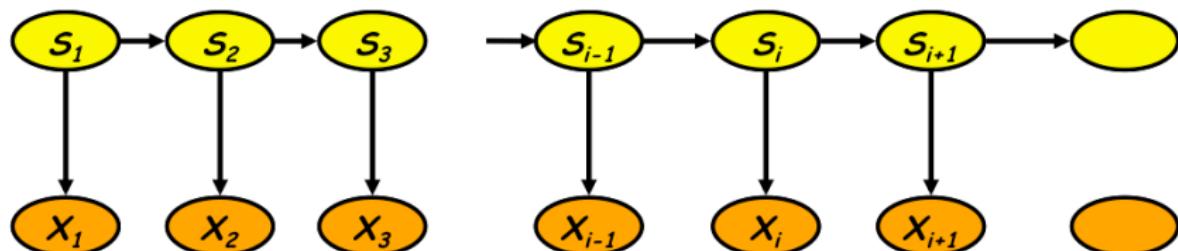
Emission probabilities: $P(X_i = b | S_i = s) = e_s(b)$

- For Markov Chains we know:

$$P(S) = P(s_1, s_2, \dots, s_L) = \prod_{i=1}^L P(s_i | s_{i-1})$$

- What is $P(s, x) = P(s_1, \dots, s_L; x_1, \dots, x_L)$?

Hidden Markov Model



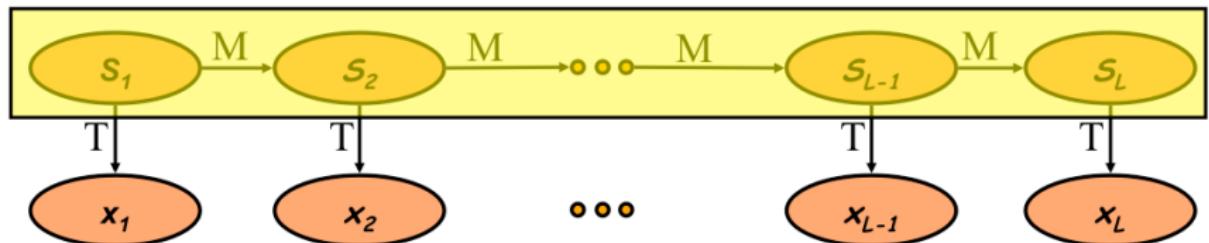
- Independence assumptions:

$$P(s_i | s_1, \dots, s_{i-1}, x_1, \dots, x_{i-1}) = P(s_i | s_{i-1})$$

$$P(x_i | s_1, \dots, s_i, x_1, \dots, x_{i-1}) = P(x_i | s_i)$$

- $P(X_i = b | S_i = s) = e_s(b)$, means that the probability of x_i depends only on the probability of s_i .

Hidden Markov Model



- The joint distribution for the full chain:

$$P(s_1, x_1, s_2, x_2, \dots, s_L, x_L) = P(s_1)P(x_1|s_1)\prod_{i=2}^L P(s_i|s_{i-1})P(x_i|s_i)$$

Why “Hidden”?

- Observers can see the emitted symbols of an HMM but have no ability to know which state the HMM is currently in.
- Thus, the goal is to infer the most likely hidden states of an HMM based on the given sequence of emitted symbols.

HMM parameters

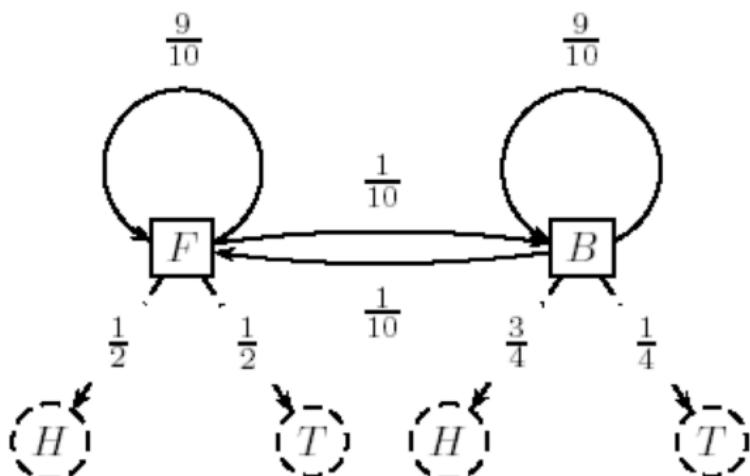
- Σ : set of emission characters.
Example.: $\Sigma = \{H, T\}$ for coin tossing
 $\Sigma = \{1, 2, 3, 4, 5, 6\}$ for dice tossing
- Q : set of hidden states, each emitting symbols from Σ .
 $Q = \{F, B\}$ for coin tossing
- $A = (a_{kl})$: a $|Q| \times |Q|$ matrix of probability of changing from state k to state l .

$$a_{FF} = 0.9, a_{FB} = 0.1, a_{BF} = 0.1, a_{BB} = 0.9$$

- $E = (e_k(b))$: a $|Q| \times |\Sigma|$ matrix of probability of emitting symbol b while being in state k .

$$e_F(H) = 1/2, e_F(T) = 1/2, e_B(T) = 1/4, e_B(H) = 3/4$$

Hidden Markov Model



HMM model for the *Fair Bet Casino Problem*

Three main problems in HMMs

- Evaluation: Given HMM parameters and observation sequence $\{x_1, x_2, \dots, x_L\}$, find probability of the observed sequence $P(x_1, x_2, \dots, x_L)$.
- Decoding: Given HMM parameters and observation sequence $\{x_1, x_2, \dots, x_L\}$, find most probable sequence of hidden states:

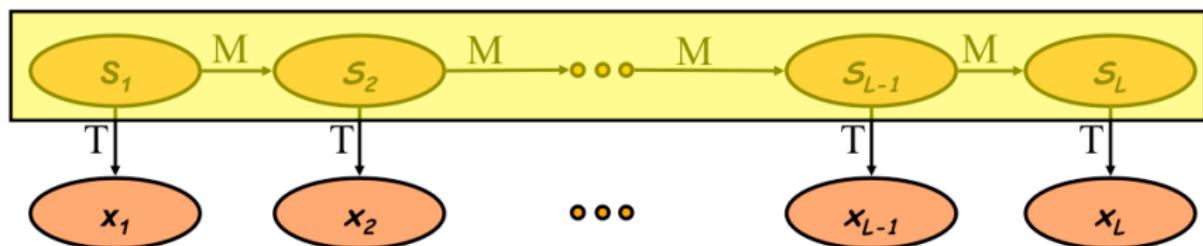
$$\arg \max P(s_1, s_2, \dots, s_L | x_1, x_2, \dots, x_L)$$

- Learning: Given HMM with unknown parameters and observation sequence, find parameters that maximize likelihood of observed data

$$\arg \max P(x_1, x_2, \dots, x_L | \Theta)$$

- Evaluation: What is the probability of the observed sequence? **Forward Algorithm**
- Decoding: What is the probability that the state s was loaded given the observed sequence? **Forward-Backward Algorithm**
What is the most likely state sequence given the observed sequence? **Viterbi Algorithm**
- Learning: Under what parameterization is the observed sequence most probable? **Baum-Welch Algorithm (EM)**

Evaluation Problem



- Given HMM parameters: transition probability matrix A and emission probability matrix E , and observation sequence $\{x_1, x_2, \dots, x_L\}$, find probability of observed sequence $P(X) = P(x_1, x_2, \dots, x_L)$.
- $P(X) = \sum_S P(X, S)$. The summation taken over all state-paths s generating x . It requires K^L terms, where K is the number of different states.
- Instead: $P(x_1, x_2, \dots, x_L) = \sum_k P(x_1, x_2, \dots, x_L, S_L = k)$

Forward algorithm for computing $P(x)$

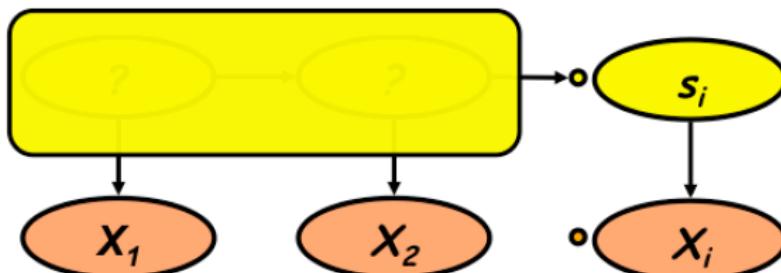
- Compute $P(x_1, x_2, \dots, x_L, S_L = k)$ recursively.
- For $i = 1, \dots, L$ and for each state l , compute:

$$f_l(i) = P(x_1, \dots, x_i; s_i = l),$$

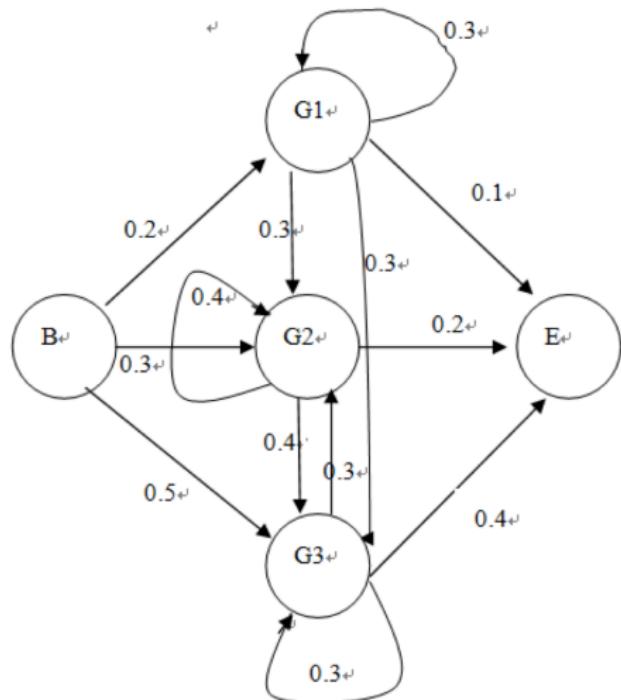
the probability of all the paths which emit (x_1, \dots, x_i) and end in state $s_i = l$.

- Use the recursive formula:

$$f_l(i) = e_l(x_i) \sum_k f_k(i-1) \cdot a_{kl}$$



Example



$$e_1(A) = 0.5, e_1(B) = 0.5$$

$$e_2(A) = 0.1, e_2(B) = 0.9$$

$$e_3(A) = 0.9, e_3(B) = 0.1$$

- Let $X = AAB$, compute $f_3(2)$.

Example

- Let $X = AAB$, compute $f_3(2)$.

$$f_1(1) = 0.5 \times 0.2 = 0.1$$

$$f_2(1) = 0.1 \times 0.3 = 0.03$$

$$f_3(1) = 0.9 \times 0.5 = 0.45$$

$$f_1(2) = 0.5 \times 0.1 \times 0.3 = 0.015$$

$$f_2(2) = 0.1 \times (0.1 \times 0.3 + 0.03 \times 0.4 + 0.45 \times 0.3) = 0.0177$$

$$f_3(2) = 0.9 \times (0.1 \times 0.3 + 0.03 \times 0.4 + 0.45 \times 0.3) = 0.1593$$

Decoding problem 1

- Problem: Given HMM parameters: transition probability matrix A and emission probability matrix E , and observation sequence $\{x_1, x_2, \dots, x_L\}$, compute for each $i = 1, \dots, L$ and for each state k the probability that $s_i = k$.

$$P(s_i|x_1, x_2, \dots, x_L) = \frac{P(s_i, x_1, x_2, \dots, x_L)}{P(x_1, x_2, \dots, x_L)}$$

- Decompose the computation:

$$\begin{aligned} P(x_1, \dots, x_L, s_i) &= P(x_1, \dots, x_i, s_i)P(x_{i+1}, \dots, x_L | x_1, \dots, x_i, s_i) \\ &= P(x_1, \dots, x_i, s_i)P(x_{i+1}, \dots, x_L | s_i) \\ &= f_{s_i}(i) \cdot b_{s_i}(i) \\ &= F(s_i)B(s_i) \end{aligned}$$

The backward algorithm

- Compute $B(s_i) = P(x_{i+1}, \dots, x_L | s_i)$ for $i = L-1, \dots, 1$ (namely, considering evidence after time slot i).
- First step, step $L-1$: Compute $B(s_{L-1})$.

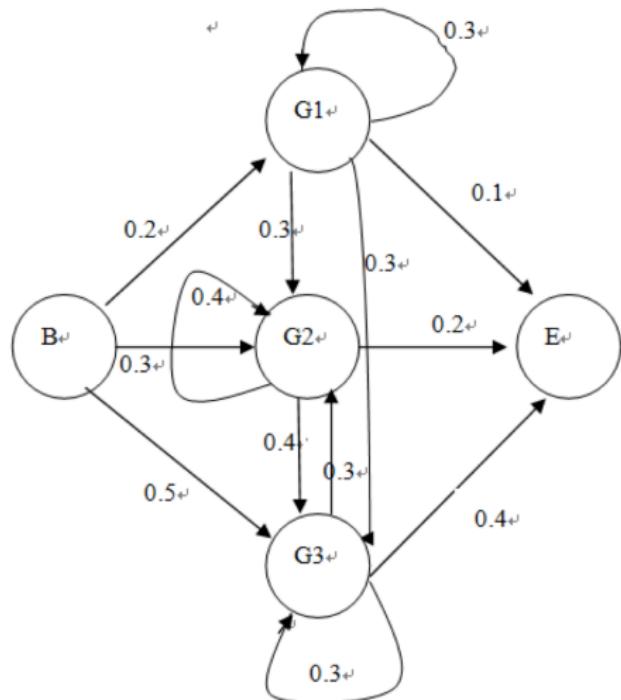
$$P(x_L | s_{L-1}) = \sum_{s_L} P(x_L, s_L | s_{L-1}) = \sum_{s_L} P(s_L | s_{L-1}) P(x_L | s_L)$$

- step i : compute $B(s_i)$ from $B(s_{i+1})$:

$$P(x_{i+1}, \dots, x_L | s_i) = \sum_{s_{i+1}} P(s_{i+1} | s_i) P(x_{i+1} | s_{i+1}) P(x_{i+2}, \dots, x_L | s_{i+1})$$

$$B(s_i) = \sum_{s_{i+1}} P(s_{i+1} | s_i) P(x_{i+1} | s_{i+1}) B(s_{i+1})$$

Example



$$e_1(A) = 0.5, e_1(B) = 0.5$$

$$e_2(A) = 0.1, e_2(B) = 0.9$$

$$e_3(A) = 0.9, e_3(B) = 0.1$$

- Let $X = AAB$, compute $b_3(2)$.

- Let $x = AAB$, compute $b_3(2)$.

$$b_1(3) = 0.1$$

$$b_2(3) = 0.2$$

$$b_3(3) = 0.4$$

$$b_1(2) = 0.3 \times 0.5 \times 0.1 + 0.3 \times 0.9 \times 0.2 + 0.3 \times 0.1 \times 0.4 = 0.081$$

$$b_2(2) = 0.4 \times 0.9 \times 0.2 + 0.4 \times 0.1 \times 0.4 = 0.088$$

$$b_3(2) = 0.3 \times 0.9 \times 0.2 + 0.3 \times 0.1 \times 0.4 = 0.066$$

$$b_1(1) = 0.3 \times 0.5 \times 0.081 + 0.3 \times 0.1 \times 0.088 + 0.3 \times 0.9 \times 0.066 = 0.03261$$

$$b_2(1) = 0.4 \times 0.1 \times 0.088 + 0.4 \times 0.9 \times 0.066 = 0.02728$$

$$b_3(1) = 0.3 \times 0.1 \times 0.088 + 0.3 \times 0.9 \times 0.066 = 0.02046$$

Most likely state vs. Most likely sequence

- Most likely state assignment at time i

$$\arg \max P(S_i = k | x_1, x_2, \dots, x_L)$$

E.g. Which die was most likely used by the casino in the third roll given the observed sequence?

- Most likely assignment of state sequence

$$\arg \max_{S_i} P(S_i, i = 1, \dots, L | x_1, x_2, \dots, x_L)$$

E.g. What was the most likely state sequence of die rolls used by the casino given the observed sequence?

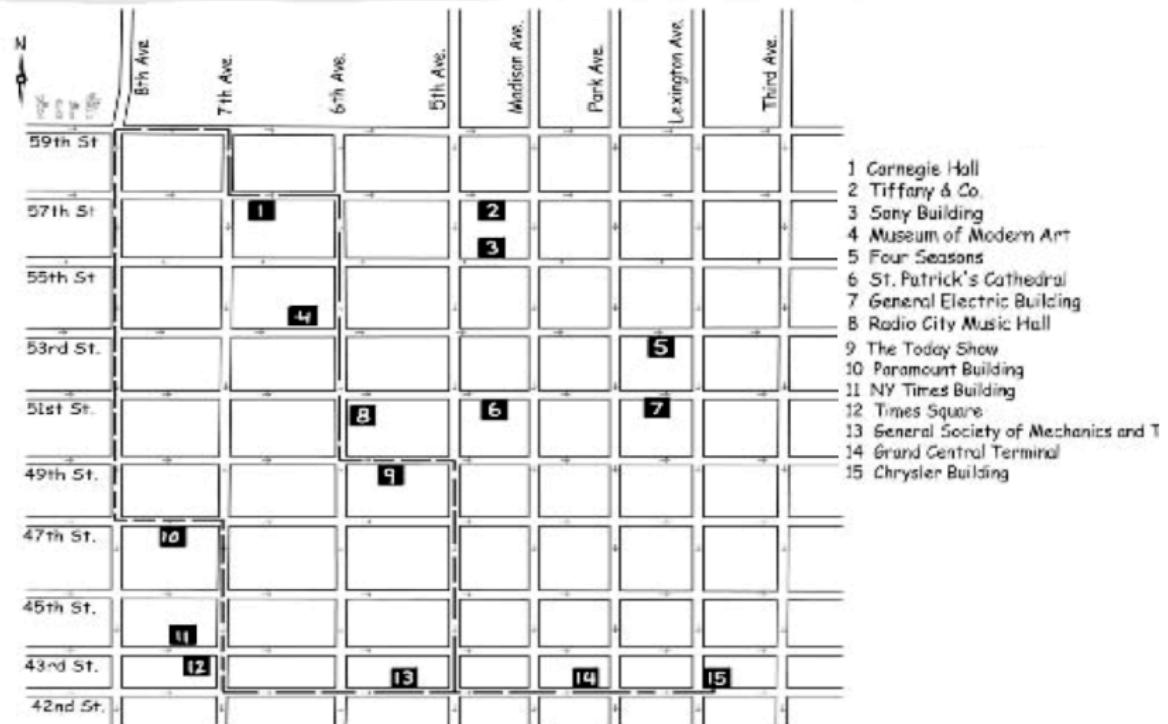
- Not the same solution !

Decoding problem 2

- Problem: Given HMM parameters: transition probability matrix A and emission probability matrix E , and observation sequence $\{x_1, x_2, \dots, x_L\}$, find most likely assignment of state sequence.
-

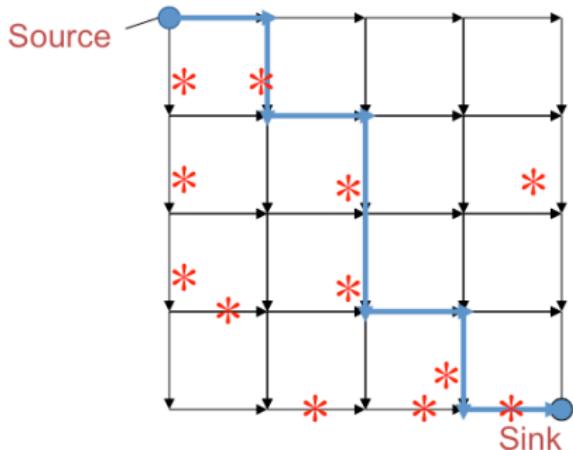
$$\begin{aligned} & \arg \max_{(s_1, s_2, \dots, s_L)} P(s_1, s_2, \dots, s_L | x_1, x_2, \dots, x_L) \\ &= \arg \max_{(s_1, s_2, \dots, s_L)} P(s_1, s_2, \dots, s_L, x_1, x_2, \dots, x_L) \\ &= \arg \max_k \max_{(s_1, s_2, \dots, s_{L-1})} P(s_1, s_2, \dots, s_{L-1}, s_L = k, x_1, x_2, \dots, x_L) \\ &= \arg \max_k v_k(L) \end{aligned}$$

Manhattan Tourist Problem (MTP)



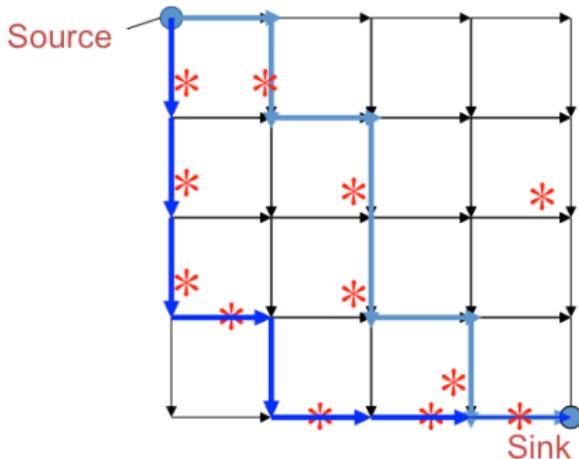
Manhattan Tourist Problem (MTP)

Imagine seeking a path (from source to sink) to travel (only eastward and southward) with the most number of attractions (*) in the Manhattan grid

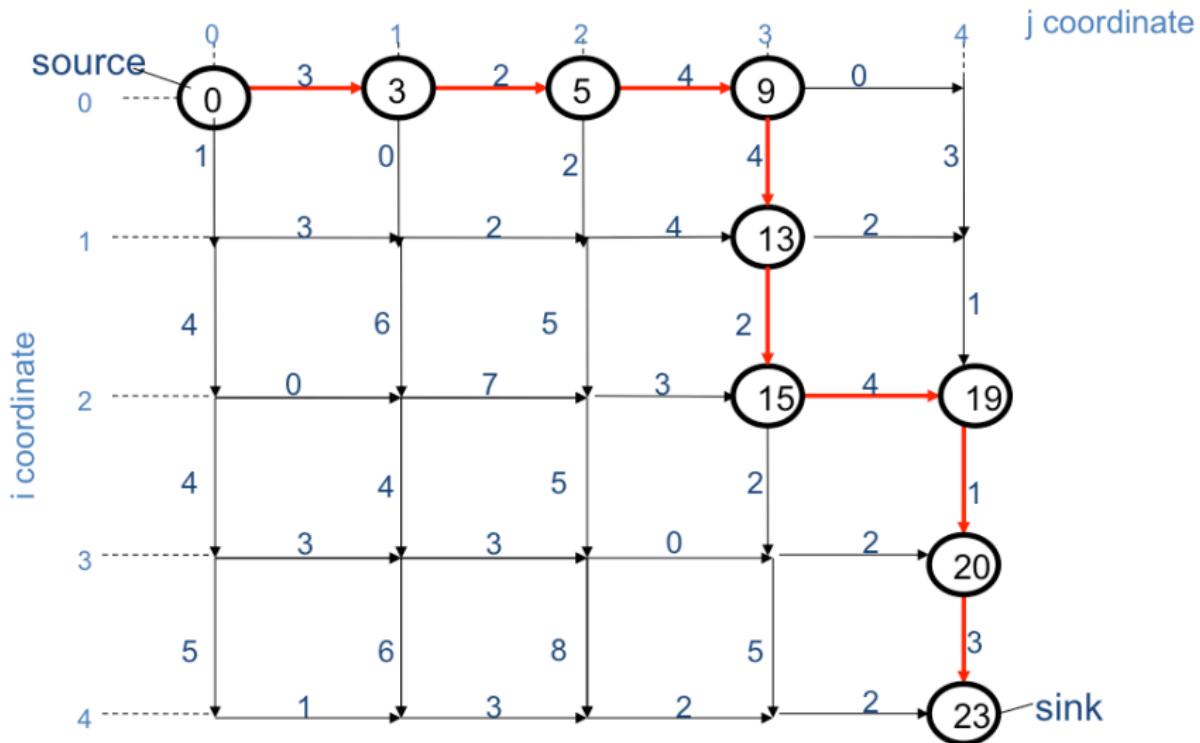


Manhattan Tourist Problem (MTP)

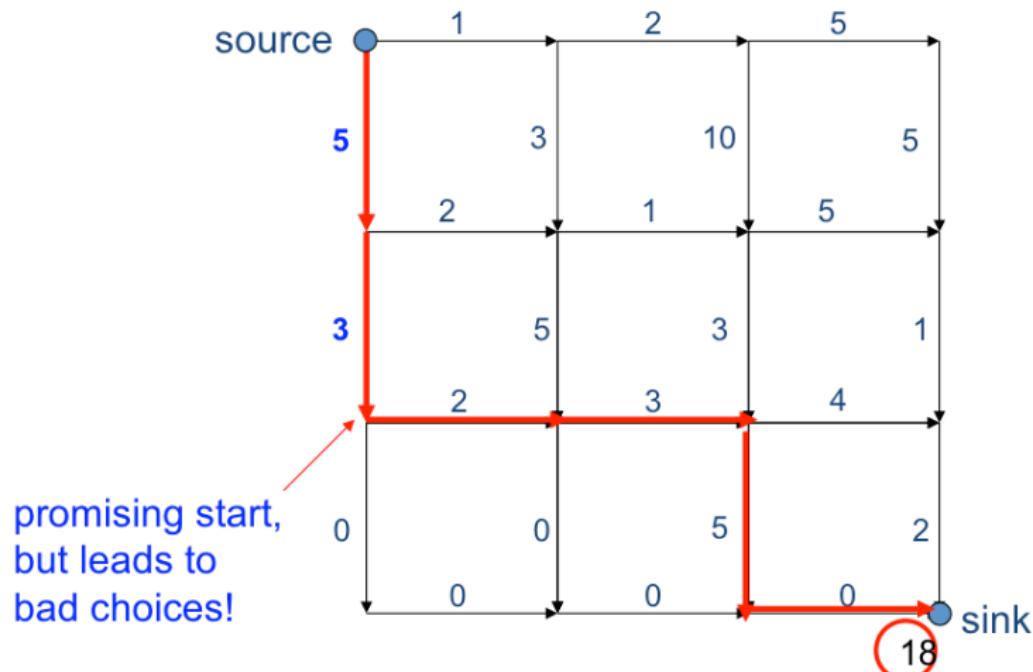
Imagine seeking a path (from source to sink) to travel (only eastward and southward) with the most number of attractions (*) in the Manhattan grid



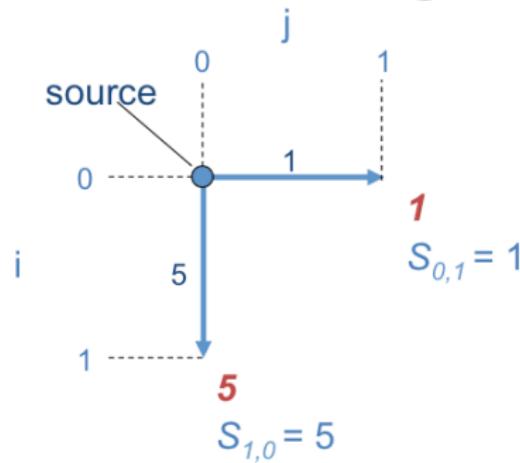
MTP: An Example



MTP: Greedy Algorithm Is Not Optimal

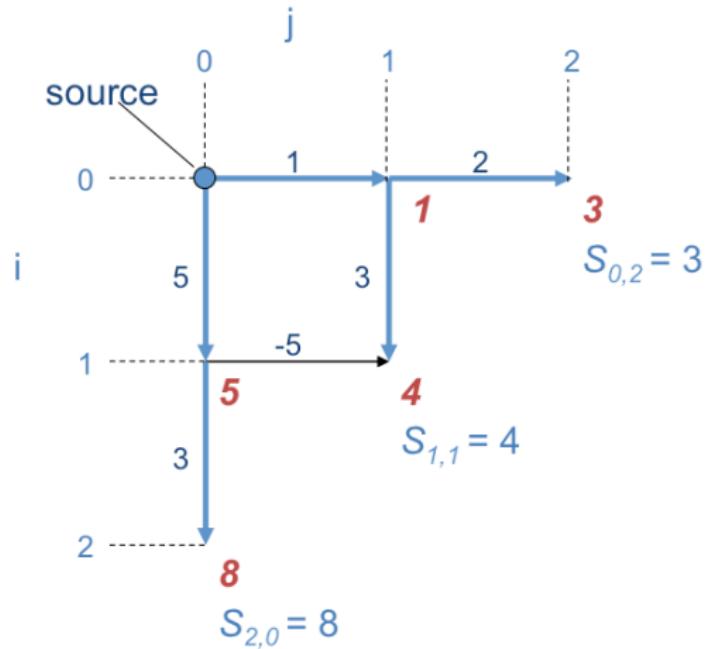


MTP: Dynamic Programming

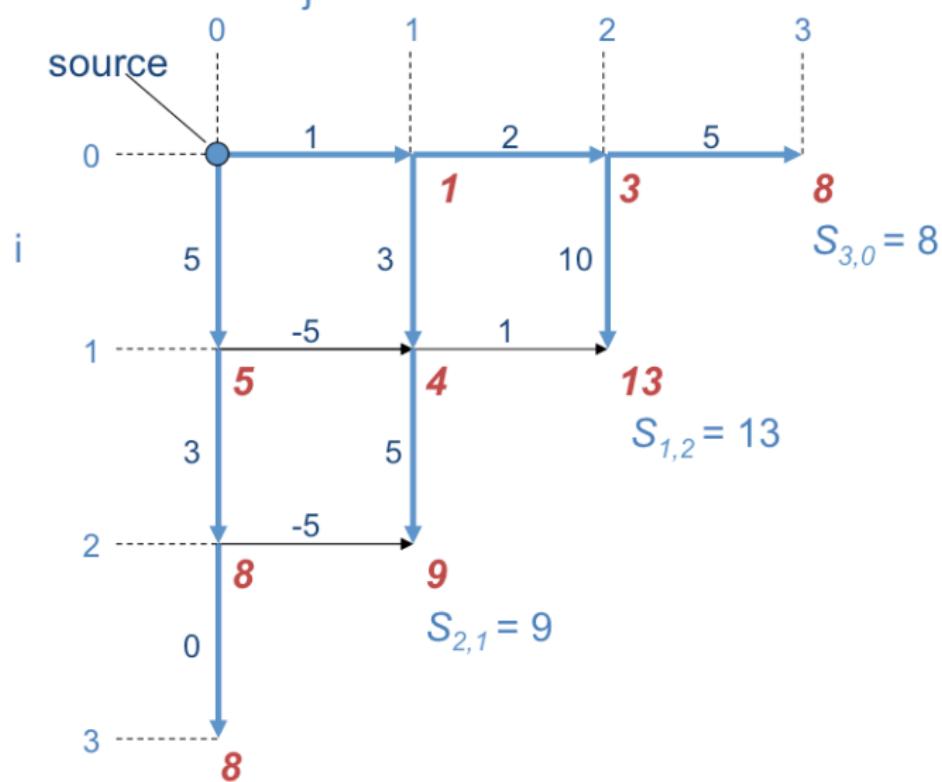


- Calculate optimal path score for each vertex in the graph
- Each vertex's score is the maximum of the prior vertices score plus the weight of the respective edge in between

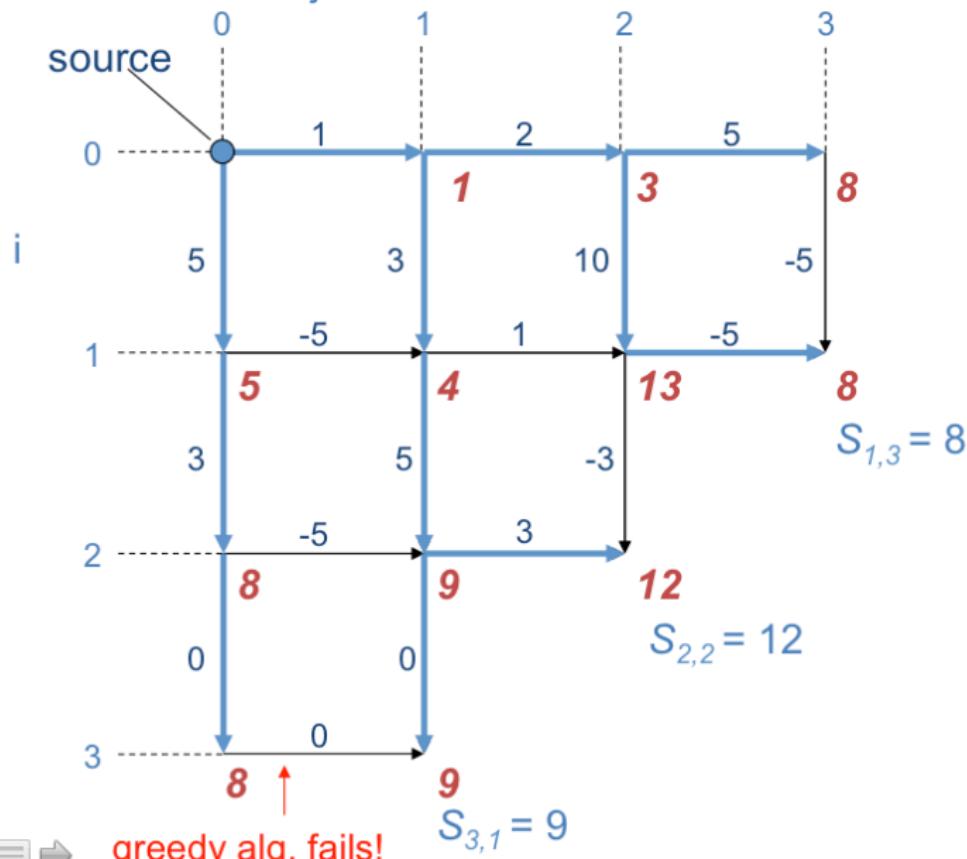
MTP: Dynamic Programming (cont'd)



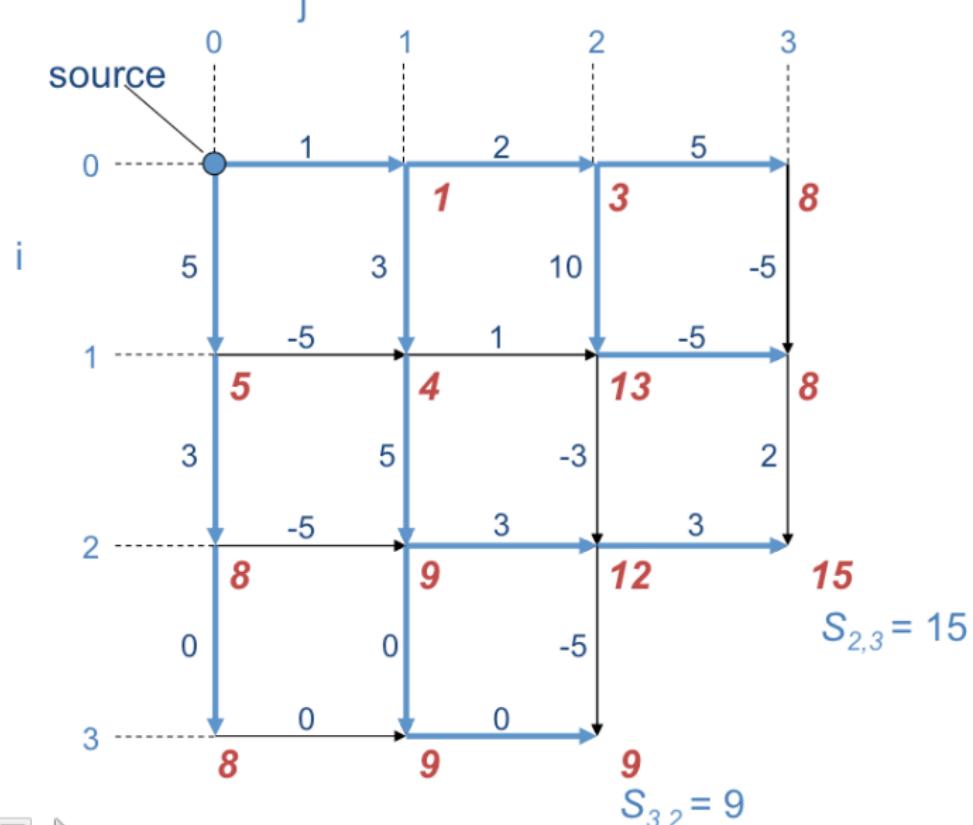
MTP: Dynamic Programming



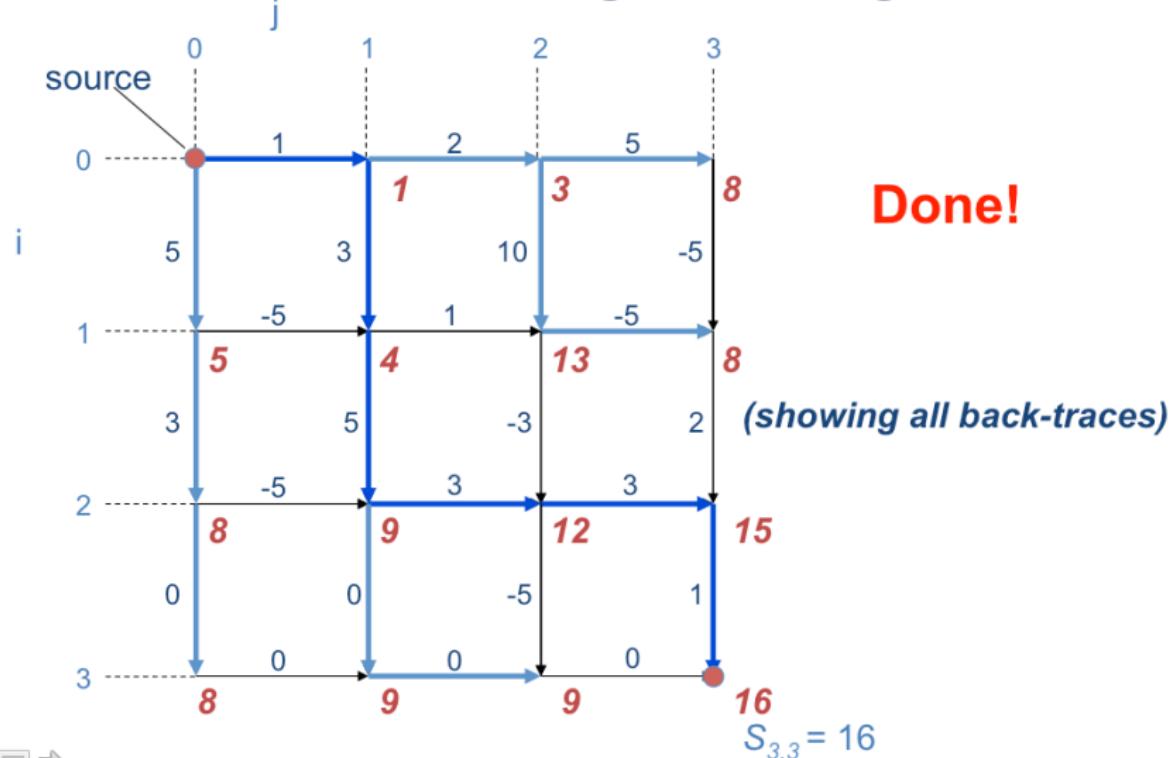
MTP: Dynamic Programming



MTP: Dynamic Programming



MTP: Dynamic Programming



MTP: Recurrence

Computing the score for a point (i,j) by the recurrence relation:

$$s_{i,j} = \max \left\{ \begin{array}{l} s_{i-1,j} + \text{weight of the edge between } (i-1, j) \text{ and } (i, j) \\ s_{i,j-1} + \text{weight of the edge between } (i, j-1) \text{ and } (i, j) \end{array} \right.$$

The running time is $n \times m$ for a n by m grid
(n = # of rows, m = # of columns)

Viterbi algorithm

- Compute probability $v_k(i)$ recursively over i .
-

$$\begin{aligned}v_k(i) &= \max_{s_1, s_2, \dots, s_{i-1}} P(s_i = k, s_1, s_2, \dots, s_{i-1}, x_1, x_2, \dots, x_i) \\&= P(x_i | s_i = k) \max_l P(s_i = k | s_{i-1} = l) v_l(i-1)\end{aligned}$$

Viterbi algorithm

- Initialization: $v_0(0) = 1, v_k(0) = 0$ for $k > 0$.
- For $i = 1$ to L do for each state l :

$$v_l(i) = e_l(x_i) \max_k v_k(i-1) a_{kl}$$

- Termination:

$$\max_{(s_1, s_2, \dots, s_L)} P(s_1, s_2, \dots, s_L, x_1, x_2, \dots, x_L) = \max_k v_k(L)$$

- Trace back: $s_L = \arg \max_k v_k(L),$
 $S_{i-1} = \arg \max_k P(s_i | s_{i-1} = k) v_k(i-1)$

Learning problem

- Given HMM with unknown parameters Θ and observation sequence $\{x_1, \dots, x_L\}$, find parameters that maximize likelihood of observed data.

$$\arg \max P(x_1, \dots, x_L | \Theta)$$

But likelihood doesn't factorize since observations are not i.i.d.

- Hidden variables: state sequence $\{s_1, s_2, \dots, s_L\}$.
- EM (Baum-Welch) Algorithm:
 - E-step: Fix parameters, find expected state assignments
 - M-step: Fix expected state assignments, update parameters

- Expectation step (E-step): Calculate the expected value of the log likelihood function, with respect to the conditional distribution of \mathbf{Z} given \mathbf{X} under the current estimate of the parameters $\theta^{(t)}$:

$$Q(\theta|\theta^{(t)}) = E_{\mathbf{Z}|\mathbf{X},\theta^{(t)}} [\log L(\theta; \mathbf{X}, \mathbf{Z})]$$

- Maximization step (M-step): Find the parameter that maximizes this quantity:

$$\theta^{(t+1)} = \arg \max_{\theta} Q(\theta|\theta^{(t)})$$

E-Step:

$$\begin{aligned} & Q(\theta|\theta^t) \\ = & \sum_{s \in S} P(s|X, \theta^t) \log P(s, X|\theta) \\ = & \sum_{s \in S} P(s|X, \theta^t) \log (\pi_{s_1} \prod_{t=1}^T a_{s_{t-1}s_t} e_{s_t}(x_t)) \\ = & \sum_{s \in S} P(s|X, \theta^t) \log \pi_{s_1} + \sum_{s \in S} P(s|X, \theta^t) \sum_{t=1}^T \log a_{s_{t-1}s_t} + \sum_{s \in S} P(s|X, \theta^t) \sum_{t=1}^T \log e_{s_t}(x_t) \end{aligned} \tag{1}$$

M-Step: The first term in Eqn. 1 becomes

$$\sum_{s \in S} P(s|X, \theta^t) \log \pi_1 = \sum_{i=1}^N P(S_1 = i|X, \theta^t) \log \pi_{1i}$$

By setting its derivative equal to zero and using the constraint $\sum_{i=1}^N \pi_{1i} = 1$, we get:

$$\pi_{1i} = P(S_1 = i|X, \theta^t)$$

The second term in Eqn.1 becomes:

$$\sum_{s \in S} P(s|X, \theta^t) \sum_{t=1}^T \log a_{s_{t-1}s_t} = \sum_{i=1}^N \sum_{j=1}^N \sum_{t=1}^T P(s_{t-1} = i, s_t = j | X, \theta^t) \log a_{ij}$$

Use a Lagrange multiplier with the constraint $\sum_{j=1}^N a_{ij} = 1$, we have

$$a_{ij} = \frac{\sum_{t=1}^T P(s_{t-1} = i, s_t = j | X, \theta^t)}{\sum_{t=1}^T P(s_{t-1} = i | X, \theta^t)}$$

$$P(s_{i-1} = k, s_i = l | X, \theta) = \frac{f_k(i-1) a_{kl} e_l(x_i) b_l(i)}{P(X|\theta)},$$

where $f_k(i-1), b_l(i)$ are forward, backward algorithms for x under θ .

$$\begin{aligned} & P(x_1, \dots, x_L, s_{i-1} = k, s_i = l) \\ &= P(x_1, \dots, x_{i-1}, s_{i-1} = k) a_{kl} e_k(x_i) P(x_{i+1}, \dots, x_L | s_i = l) \end{aligned}$$

The third term in Eqn.1 becomes

$$\sum_{s \in S} P(s|X, \theta^t) \left(\sum_{t=1}^T \log e_{s_t}(x_t) \right) = \sum_{i=1}^N \sum_{t=1}^T P(s_t = i|X, \theta^t) \log e_i(x_t)$$

Similarly, by using a Lagrange multiplier with constraint $\sum_{j=1}^L e_i(j) = 1$, we get

$$e_i(k) = \frac{\sum_{t=1}^T P(s_t = i|X, \theta^t) \delta_{s_t, v_k}}{\sum_{t=1}^T P(s_t = i|X, \theta^t)}.$$

From another point of view. Here we assume there are n observed sequences. x^j denotes the j -th sequence.

- Step 1: For each pair (k, l) , compute the expected number of state transitions from k to l :

$$\begin{aligned} A_{kl} &= \sum_{j=1}^n \frac{1}{P(x^j)} \sum_{i=1}^L P(x^j, s_{i-1} = k, s_i = l) \\ &= \sum_{j=1}^n \frac{1}{P(x^j)} \sum_{i=1}^L f_k^j(i-1) a_{kl}^j e_l^j(x_i) b_l^j(i) \end{aligned}$$

- Step 2: For each state k and each symbol b , compute the expected number of emissions of b from k :

$$E_k(b) = \sum_{j=1}^n \frac{1}{P(x^j)} \sum_{i, x_i^j = b} f_k^j(i) b_k^j(i)$$

- Step 3: Use the A_{kl} 's, $E_k(b)$'s to compute the new values of a_{kl} and $e_k(b)$. These values define Θ .

$$a_{kl} = \frac{A_{kl}}{\sum_{l'} A_{kl'}}, e_k(b) = \frac{E_k(b)}{\sum_{b'} E_k(b')}$$

- This procedure is iterated, until some convergence criterion is met.