1 Markov Chains

Given a finite set $\mathcal{Q} = \{1, 2, \ldots, I\}$ of states, called the *state space*, for some positive integer I. Let Q_t be a random variable which takes at every discrete time step t a value of $\mathcal{Q}$. A Markov chain is a sequence of random variables $Q_1, Q_2, \ldots$ and is described in terms of its *transition probabilities*

$$a_{ij} = P(Q_{t+1} = j \,|\, Q_t = i), \quad i, j \in \mathcal{Q}. \tag{1}$$

A Markov chain obeys the property that given the present state, the future and past states are independent. That is, the next state Q_{t+1} depends only on the present state Q_t. More formally, a *first-order* Markov chain is given by

$$\begin{aligned} P(Q_{t+1} = j \,|\, Q_t = i, Q_{t-1} = i_{t-1}, \ldots, Q_1 = i_1) &= P(Q_{t+1} = j \,|\, Q_t = i) &(2)\\ &= a_{ij}, &(3)\end{aligned}$$

for all times t, all states $i, j \in \mathcal{Q}$, and all possible state sequences $i_1, i_2, \ldots, i_{t-1}$. The joint distribution for a sequence of states has the form

$$P(Q_1 = i_1, Q_2 = i_2, \ldots, Q_T = i_T) = P(Q_1 = i_1) \prod_{t=2}^{T} P(Q_t = i_t \,|\, Q_{t-1} = i_{t-1}). \tag{4}$$

A *second-order* Markov chain is analogous given by

$$\begin{aligned} &P(Q_{t+1} = k \,|\, Q_t = j, Q_{t-1} = i, Q_{t-2} = i_{t-2}, \ldots, Q_1 = i_1) = \\ &P(Q_{t+1} = k \,|\, Q_t = j, Q_{t-1} = i), &(5)\end{aligned}$$

for all times t, all states $i, j, k \in \mathcal{Q}$, and all possible state sequences $i_1, i_2, \ldots, i_{t-2}$. The joint distribution for a sequence of states has the form

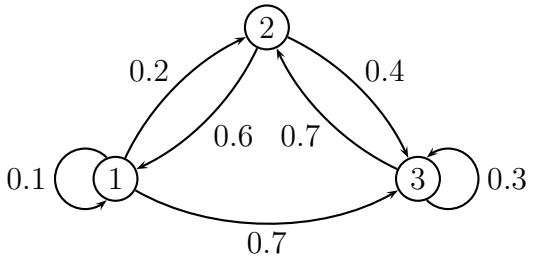
$$\begin{aligned} P(Q_1 = i_1, Q_2 = i_2, \ldots, Q_T = i_T) = P(Q_1 = i_1) P(Q_2 = i_2 \,|\, Q_1 = i_1) \times \\ \prod_{t=3}^{T} P(Q_t = i_t \,|\, Q_{t-1} = i_{t-1}, Q_{t-2} = i_{t-2}). &(6)\end{aligned}$$

Higher order Markov chains can be defined according to the same principle. It is interesting to observe, that an l-order Markov chain can be defined as a first-order Markov chain. However $O(I^l)$ states are required.

In summary, a Markov chain is specified by an $I \times I$ transition matrix $\mathbf{A} = \{a_{ij}\}$ which is row stochastic ($\sum_{j=1}^{I} a_{ij} = 1$) and an initial state distribution $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_I)$, where $\pi_i = P(Q_1 = i)$ and $\sum_{i=1}^{I} \pi_i = 1$. The parameters required to specify a Markov chain are denoted as $\Theta = (\boldsymbol{\pi}, \mathbf{A})$ and summarized in Table 1.

$\mathcal{Q} = \{1, 2, \ldots, I\}$	set of states, where I is the number of states.
$\mathbf{A} = \{a_{ij}\}$	$I \times I$ state transition matrix.
$\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_I)$	initial state distribution.
$\mathbf{s} = i_1, i_2, \ldots, i_T$	state sequence of length T.
$\Theta = (\boldsymbol{\pi}, \mathbf{A})$	model parameters.

Table 1: Markov chain parameters.



$$\mathbf{A} = \begin{bmatrix} 0.1 & 0.2 & 0.7 \\ 0.6 & 0 & 0.4 \\ 0 & 0.7 & 0.3 \end{bmatrix}$$

Figure 1: Markov chain and the corresponding transition matrix $\mathbf{A}$.

2 Probability of Path

Given a sequence $\mathbf{s} = i_1, i_2, \ldots, i_T$ of states, the probability is

$$P(\mathbf{s}) = P(Q_1 = i_1, Q_2 = i_2, \ldots, Q_T = i_T) = P(Q_1 = i_1) \cdot a_{i_1 i_2} \cdot a_{i_2 i_3} \cdot \ldots \cdot a_{i_{T-1} i_T}.$$

If the initial state $Q_1 = i_1$ is known, we have

$$P(Q_2 = i_2, Q_3 = i_3, \ldots, Q_T = i_T \,|\, Q_1 = i_1) = a_{i_1 i_2} \cdot a_{i_2 i_3} \cdot \ldots \cdot a_{i_{T-1} i_T}.$$

Example 2.1

$$\begin{aligned} P(Q_1 = 2, Q_2 = 3, Q_3 = 3, Q_4 = 2, Q_5 = 1) &= P(Q_1 = 2) \cdot a_{23}\, a_{33}\, a_{32}\, a_{21} \\ &= P(Q_1 = 2) \cdot 0.4 \cdot 0.3 \cdot 0.7 \cdot 0.6. \end{aligned}$$

$$\begin{aligned} P(Q_2 = 3, Q_3 = 3, Q_4 = 2, Q_5 = 1 \,|\, Q_1 = 2) &= a_{23}\, a_{33}\, a_{32}\, a_{21} \\ &= 0.4 \cdot 0.3 \cdot 0.7 \cdot 0.6. \end{aligned}$$

In general the probability of a sequence $\mathbf{s}$ given the model parameters Θ is

$$P(\mathbf{s} \,|\, \Theta) = P(Q_1 = i_1) \prod_{t=2}^{T} P(Q_t = i_t \,|\, Q_{t-1} = i_{t-1}). \tag{7}$$

The beauty of Markov chains is the calculation of the probability of future states, conditioned on the current state. This probability is captured by the t-step transition probability and defined by

$$r_{ij}(t) = P(Q_t = j \,|\, Q_1 = i). \tag{8}$$

Expression (8) is the probability that the state after t time steps will be j, given that the current state is i. It can be calculated by the following recursive formula which was independently derived by Chapman and Kolmogorov

$$r_{ij}(t) = \begin{cases} \sum_{m=1}^{M} r_{im}(t-1) a_{mj} & \text{for } t > 1, \text{ and all } i, j \\ a_{ij} & \text{for } r_{ij}(1). \end{cases} \tag{9}$$

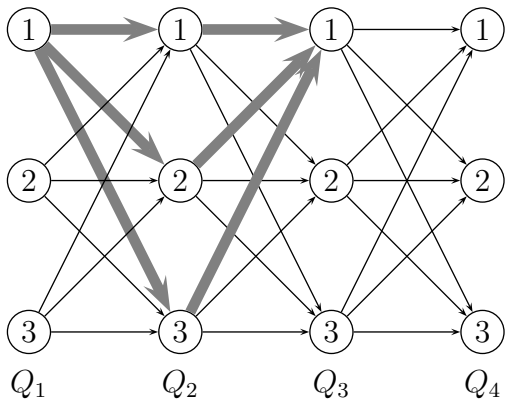
Example 2.2
We consider the Markov model in Figure (1) and calculate $r_{12}(3) = P(Q_4 = 2 \,|\, Q_1 = 1)$, that is, the

probability that after 3 time steps the final state is 2, given that the current state is 1. All possible paths from state 1 to 2 over 3 time steps are

Q_1	Q_2	Q_3	Q_4
1	1	1	3
1	1	2	3
1	1	3	3
1	2	1	3
1	2	2	3
1	2	3	3
1	3	1	3
1	3	2	3
1	3	3	3

$$r_{12}(3) = \sum_{m=1}^{M=3} r_{1m}a_{m3} = r_{11}(2)a_{13} + r_{12}(2)a_{23} + r_{13}(2)a_{33} = P(Q_4 = 2 \,|\, Q_1 = 1). \tag{10}$$

In Figure 2 all paths from state 1 to 1 over 2 time steps are depicted. This corresponds to the sub-term $r_{11}(2)$ in Equation (10).



$$\begin{aligned} r_{11}(2) &= r_{11}(1)a_{11} + r_{12}(1)a_{21} + r_{13}(1)a_{31} \\ &= a_{11}a_{11} + a_{12}a_{21} + a_{13}a_{31} \\ &= \text{ paths gray arrows} \end{aligned}$$

Figure 2: Recursive formula for sub-term $r_{11}(2)$ and corresponding paths.

3 Inference

Given a sequence $\mathbf{s} = i_1, i_2, \ldots, i_T$ and number of states I. Our goal is to infer $\Theta = (\boldsymbol{\pi}, \mathbf{A})$ with the *maximum likelihood estimation*. The likelihood of $\mathbf{s}$

$$l(\Theta \,|\, \mathbf{s}) = P(Q_1 = i_1) \prod_{t=2}^{T} P(Q_t = i_t \,|\, Q_{t-1} = i_{t-1}) \tag{11}$$

$$= \pi_{i_1} \prod_{t=2}^{T} a_{i_{t-1}\, i_t}. \tag{12}$$

Denote n_{ij} the number of times state i is followed by j in $\mathbf{s}$, (11) can be rewritten as follows

$$l(\Theta \,|\, \mathbf{s}) = \pi_i \prod_{i=1}^{I} \prod_{j=1}^{I} a_{ij}^{n_{ij}} \tag{13}$$

The log-likelihood of $\mathbf{s}$ is

$$\mathcal{L}(\Theta \,|\, \mathbf{s}) \;=\; \log\left[\pi_i \prod_{i=1}^{I}\prod_{j=1}^{I} a_{ij}^{n_{ij}}\right] = \log \pi_i + \sum_{i=1}^{I}\sum_{j=1}^{I} \log a_{ij}^{n_{ij}}. \tag{14}$$

Maximizing (14) subject to the constrains that each row in $\mathbf{A}$ and vector $\boldsymbol{\pi}$ sums to one is

$$\text{maximize} \quad \mathcal{L}(\Theta \,|\, \mathbf{s}) \tag{15}$$

$$\text{subject to} \quad \sum_{j=1}^{I} a_{ij} = 1 \qquad \text{for all } i = 1, \ldots, I \tag{16}$$

$$\sum_{i=1}^{I} \pi_i = 1. \tag{17}$$

The optimization problem can be solved with the method of Lagrange multipliers. Since we have in total $I+1$ constraints, namely $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_I)$ and ξ we obtain the form

$$L(\boldsymbol{\pi}, \mathbf{A}, \boldsymbol{\lambda}, \xi) = \log \pi_i + \sum_{i=1}^{I}\sum_{j=1}^{I} \log a_{ij}^{n_{ij}} + \sum_{i=1}^{I} \lambda_i \left(1 - \sum_{j=1}^{I} a_{ij}\right) + \xi\left(1 - \sum_{i=1}^{I} \pi_i\right). \tag{18}$$

Differentiating component-wise and setting the derivatives to zero yields

$$\frac{\partial L}{\partial \pi_i} \;=\; \frac{1}{\pi_i} - \xi = 0 \tag{19}$$

$$\frac{\partial L}{\partial a_{ij}} \;=\; \frac{n_{ij}}{a_{ij}} - \lambda_i = 0 \tag{20}$$

$$\frac{\partial L}{\partial \lambda_i} \;=\; 1 - \sum_{j=1}^{I} a_{ij} = 0 \tag{21}$$

$$\frac{\partial L}{\partial \xi} \;=\; 1 - \sum_{i=1}^{I} \pi_i = 0. \tag{22}$$

$$\tag{23}$$

From (20) and (21) we get

$$\frac{n_{ij}}{\lambda_i} = a_{ij} \quad \text{and} \quad \sum_{j=1}^{I} \frac{n_{ij}}{\lambda_i} = 1 \Leftrightarrow \sum_{j=1}^{I} n_{ij} = \lambda_i. \tag{24}$$

From (19) and (22) we get

$$\frac{1}{\xi} = \pi_i \quad \text{and} \quad 1 = \sum_{i=1}^{I} \frac{1}{\xi} \Leftrightarrow \xi = I \tag{25}$$

and finally the result

$$\hat{a}_{ij} = \frac{n_{ij}}{\sum_{j=1}^{I} n_{ij}} \quad \text{and} \quad \hat{\pi}_i = \frac{1}{I}. \tag{26}$$

Given a sample $\mathcal{S} = \{\mathbf{s}_n\}_{n=1}^{N}$. Now, n_{ij} denotes the number of times state i is followed by j in all sequences of $\mathcal{S}$ and denote n_i the number of times a sequence starting in state i in $\mathcal{S}$. The likelihood of $\mathcal{S}$ is

$$l(\Theta \,|\, \mathcal{S}) = \pi_i^{n_i} \prod_{i=1}^{I}\prod_{j=1}^{I} a_{ij}^{n_{ij}} \tag{27}$$

and consequently the log-likelihood results in

$$\mathcal{L}(\Theta \,|\, \mathcal{S}) = \log \pi_i^{n_i} \sum_{i=1}^{I} \sum_{j=1}^{I} \log a_{ij}^{n_{ij}}. \tag{28}$$

Following the same approach as before, yields the maximum likelihood estimation result

$$\hat{a}_{ij} = \frac{n_{ij}}{\sum_{j=1}^{I} n_{ij}} \quad \text{and} \quad \hat{\pi}_i = \frac{n_i}{\sum_{i=1}^{I} n_i}. \tag{29}$$

4 Hidden Markov Model

Hidden Markov Models are probabilistic generative models frequently applied in fields such as speech recognition [4] and biological sequence analysis [3]. A Hidden Markov Model (HMM) is a Markov chain with the extension, that each state emits according to a multinomial distribution a symbol. A HMM is specified by parameters

$\mathcal{Q} = \{1, 2, \ldots, I\}$	set of states, where I is the number of states.
$\mathcal{O} = \{1, 2, \ldots, M\}$	set of symbols, where M is the number of symbols.
$\mathbf{A} = \{a_{ij}\}$	$I \times I$ state transition probability matrix.
$\mathbf{B} = \{b_{im}\}$	$I \times M$ emission probability matrix.
$\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_I)$	initial state distribution.
$\mathbf{s} = i_1, i_2, \ldots, i_T$	state sequence of length T.
$\mathbf{o} = o_1, o_2, \ldots, o_T$	output sequence of length T.
$\Theta = (\boldsymbol{\pi}, \mathbf{A}, \mathbf{B})$	model parameters.

Table 2: Hidden Markov model parameters.

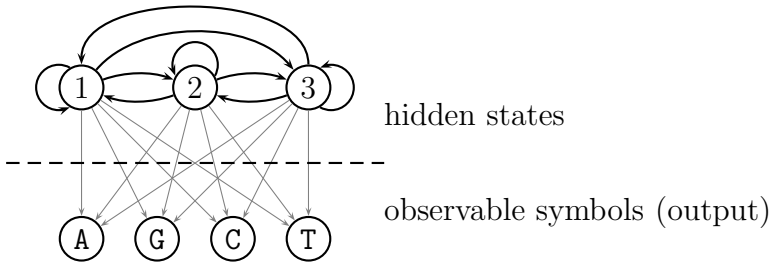


Figure 3: A Hidden Markov Model with three hidden states $\mathcal{Q} = \{1, 2, 3\}$ and four symbols

4.1 Probability of an observation sequence

Given a HMM with parameter Θ and an observation sequence $\mathbf{o}$. We wish to evaluate $P(\mathbf{o} \,|\, \Theta)$. The probability for $\mathbf{o}$ given state sequence $\mathbf{q}$ is

$$P(\mathbf{o} \,|\, \mathbf{q}, \Theta) = \prod_{t=1}^{T} P(O_t = o_t \,|\, Q_t = q_t, \Theta) \tag{30}$$

References

[1] Ethem Alpaydin. *Introduction to Machine Learning*. MIT Press, 2004.

[2] Dimitri P. Bertsekas and John N. Tsitsiklis. *Introduction to Probability*. Athena Scientific, 2002.

[3] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

[4] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.