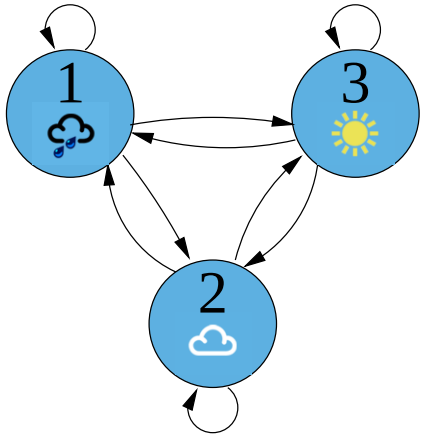
UNIVERSITY OF SURREY
Centre for Vision Speech & Signal Processing
University of Surrey, Guildford GU2 7XH.

HMM part 1

Dr Philip Jackson

- Probability fundamentals
- Markov models
- State topology diagrams
- Hidden Markov models
  - Likelihood calculation
  - Recognition & training



Summary of Dynamic Time Warping

The DTW approach allows efficient computation with limited flexibility in the alignment. It treats templates as deterministic with residual noise.

Problems:

1. How much flexibility should we allow?
2. How should we penalise any warping?
3. How do we determine a fair distance metric?
4. How many templates should we register?
5. How do we select the best ones?

Solution:

- Develop an inference framework to build templates based on the statistics of our data.

Characteristics of the desired model

1. evolution of sequence should not be deterministic
2. observations are coloured depending on class
3. cannot directly observe class
4. stochastic sequence + stochastic observations

Applications:

- automatic speech recognition
- optical character recognition
- protein and DNA sequencing
- speech synthesis
- noise-robust data transmission
- crytoanalysis
- machine translation
- image classification, etc.

Probability fundamentals

- Normalisation
- Independent events
- Dependent events
- Bayes' theorem
- Marginalisation

Normalisation

Discrete: probability of all possibilities sums to one:

$$\sum_{\text{all } X} P(X) = 1 \tag{1}$$

Continuous: integral over the entire probabilty density function (pdf) comes to one:

$$\int_{-\infty}^{\infty} p(x)\, dx = 1 \tag{2}$$

Joint probability

The joint probability that two **independent** events occur is the product of their individual probabilities:

$$P(A, B) = P(A)\, P(B) \tag{3}$$

Conditional probability

If two events are **dependent**, we need to determine their conditional probabilities. The joint probability is now

$$P(A, B) = P(A)\, P(B|A) \tag{4}$$

where $P(B|A)$ is the probability of event B given that A occurred; conversely, taking the events the other way

$$P(A, B) = P(A|B)\, P(B) \tag{5}$$

$P(A,B)$	A	$\bar{A}$	
B	0.1	0.3	0.4
$\bar{B}$	0.4	0.2	0.6
	0.5	0.5	1.0

These expressions can be rearranged to yield the conditional probabilities. Also, we can combine them to obtain the theorem proposed by Rev. Thomas Bayes (C.18th).

Bayes' theorem

Equating the RHS of eqs. 4 and 5 gives

$$P(B|A) = \frac{P(A|B)\,P(B)}{P(A)} \tag{6}$$

For example, in a word recognition application we have

$$P(w|\mathcal{O}) = \frac{p(\mathcal{O}|w)\,P(w)}{p(\mathcal{O})} \tag{7}$$

which can be interpreted as

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} \tag{8}$$

posterior probability gives basis for Bayesian inference,
likelihood describes how likely are data for given class,
prior incorporates other knowledge (e.g., language model),
evidence normalises and is often discarded (as it is the same for all classes).

Marginalisation

Discrete: probability of event B, which depends on A, is the sum over A of all joint probabilities:

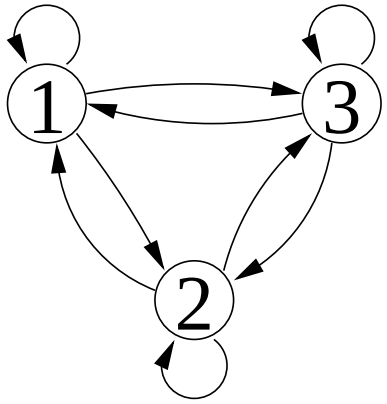
$$P(B) = \sum_{\text{all } A} P(A, B) = \sum_{\text{all } A} P(B|A)\, P(A) \tag{9}$$

Continuous: similarly, the nuisance factor x can be eliminated from its joint pdf with y:

$$p(y) = \int_{-\infty}^{\infty} p(x, y)\, dx = \int_{-\infty}^{\infty} p(y|x)p(x)\, dx \tag{10}$$

Introduction to Markov models

We can model stochastic sequences using a Markov chain, e.g., the state topology of an ergodic Markov model:



For 1st-order Markov chains, probability of state occupation depends only on the previous step (Rabiner, 1989):

$$P(x_t = j | x_{t-1} = i, x_{t-2} = h, \ldots) \approx P(x_t = j | x_{t-1} = i) \quad (11)$$

So, if we assume the RHS of eq. 11 is independent of time, we can write the **state-transition probabilities**

$$a_{ij} = P(x_t = j | x_{t-1} = i), \quad 1 \leq i, j \leq N \quad (12)$$

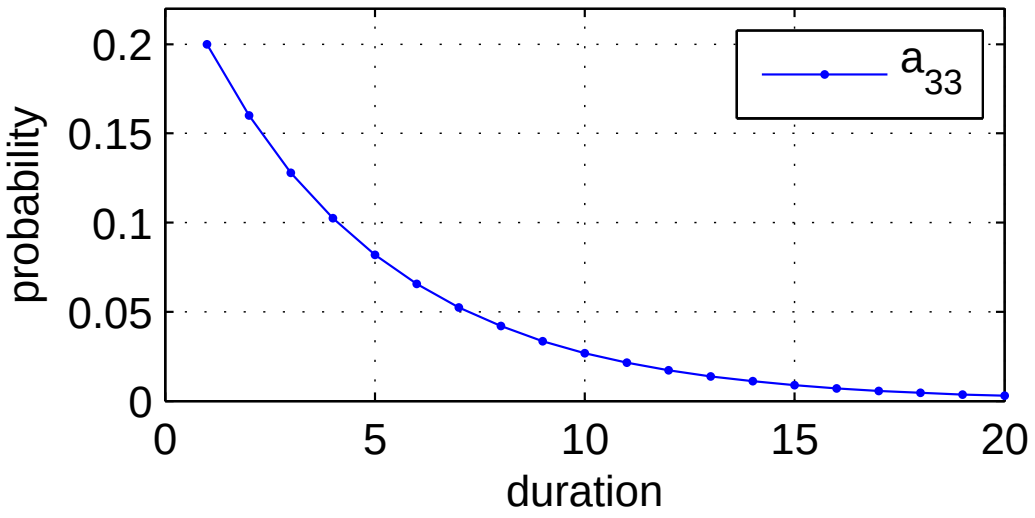
with the properties

$$a_{ij} \geq 0 \qquad \text{and} \qquad \sum_{j=1}^{N} a_{ij} = 1 \qquad \forall\ i, j \in 1..N$$

State duration characteristics

As a consequence of the first-order Markov model, the probability of occupying a state for a given duration, τ, decays exponentially:

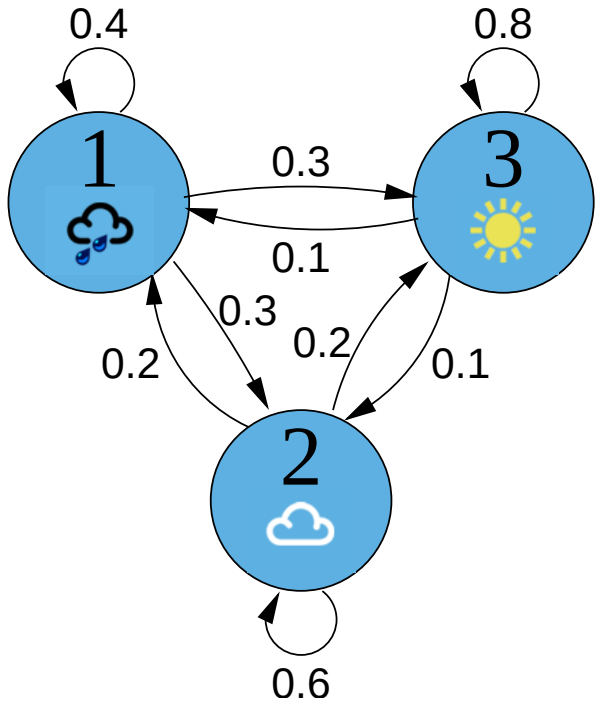
$$p(X|x_1 = i, \mathcal{M}) = (a_{ii})^{\tau-1}\ (1 - a_{ii}) \tag{13}$$



Weather prediction example

Let us represent the state of the weather by a 1st-order, ergodic Markov model, $\mathcal{M}$:

state 1: raining
state 2: cloudy
state 3: sunny



with state-transition probabilities expressed in matrix form:

$$A = \{a_{ij}\} = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix} \tag{14}$$

Weather predictor probability calculation

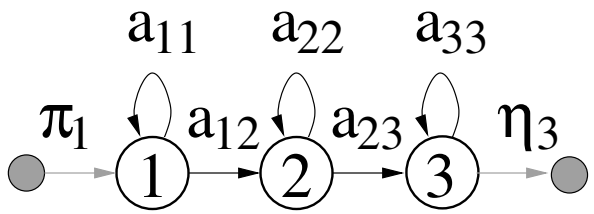
Given today's weather what is the probability of directly observing the sequence of weather states "rain-sun-sun" with model $\mathcal{M}$?

$$A = \begin{array}{c} \\ \text{rain} \\ \text{cloud} \\ \text{sun} \end{array} \begin{array}{c} \begin{array}{ccc} \text{rain} & \text{cloud} & \text{sun} \end{array} \\ \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix} \end{array}$$

$$\begin{aligned} P(X|\mathcal{M}) &= P(X = \{1,3,3\}|\mathcal{M}) \\ &= P(x_1 = \text{rain}|\text{today}) \times P(x_2 = \text{sun}|x_1 = \text{rain}) \\ &\quad \times P(x_3 = \text{sun}|x_2 = \text{sun}) \\ &= a_{\ 1} a_{13} a_{33} \\ &= \quad \times 0.3 \times 0.8 \\ &= \end{aligned}$$

Start and end of a state sequence

Null states deal with the start and end of sequences, as in the state topology of this left-right Markov model:



Entry probabilities at t=1 for each state i are defined

$$\pi_i = P(x_1 = i) \qquad 1 \leq i \leq N \tag{15}$$

with the properties $\pi_i \geq 0$, and $\sum_{i=1}^{N} \pi_i = 1$ for $i \in 1..N$

Exit probabilities at t=T are similarly defined

$$\eta_i = P(x_T = i) \qquad 1 \leq i \leq N \tag{16}$$

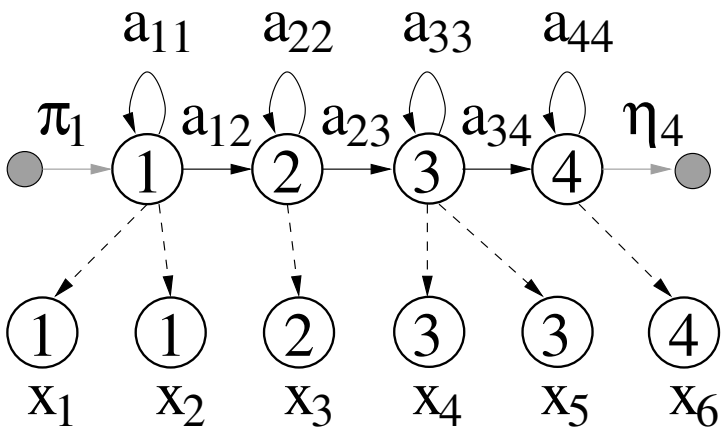
with properties $\eta_i \geq 0$, and $\eta_i + \sum_{j=1}^{N} a_{ij} = 1$ for $i \in 1..N$

Parameters of the Markov Model, $\mathcal{M}$

State transition probabilities,

$A = \{\pi_j, a_{ij}, \eta_i\} = \{P(x_t = j | x_{t-1} = i)\}$ for $1 \leq i, j \leq N$

where N is the number of states

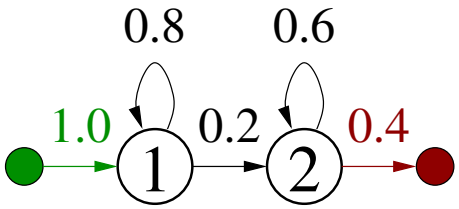


producing a sequence

$X = \{1, 1, 2, 3, 3, 4\}$

Example: probability of MM state sequence

Consider the state topology



state transition probabilities

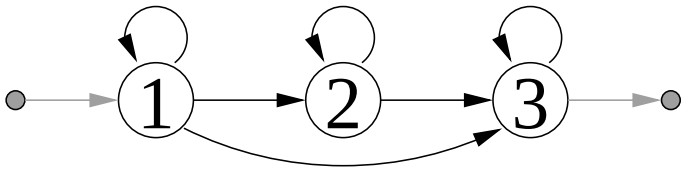
$$A = \left[\begin{array}{c|cc|c} 0 & 1 & 0 & 0 \\ \hline 0 & 0.8 & 0.2 & 0 \\ 0 & 0 & 0.6 & 0.4 \\ \hline 0 & 0 & 0 & 0 \end{array}\right]$$

The probability of state sequence $X = \{1, 2, 2\}$ is

$$\begin{aligned} P(X|\mathcal{M}) &= \pi_1 \, a_{12} \, a_{22} \, \eta_2 \\ &= 1 \times 0.2 \times 0.6 \times 0.4 \\ &= 0.048 \end{aligned}$$

Summary of Markov models

State topology diagram:



Entry probabilities $\boldsymbol{\pi} = \{\pi_i\} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$ and exit probabilities $\boldsymbol{\eta} = \{\eta_i\} = \begin{bmatrix} 0 & 0 & 0.2 \end{bmatrix}^{\mathsf{T}}$ are combined with state transition probabilities in complete A matrix:

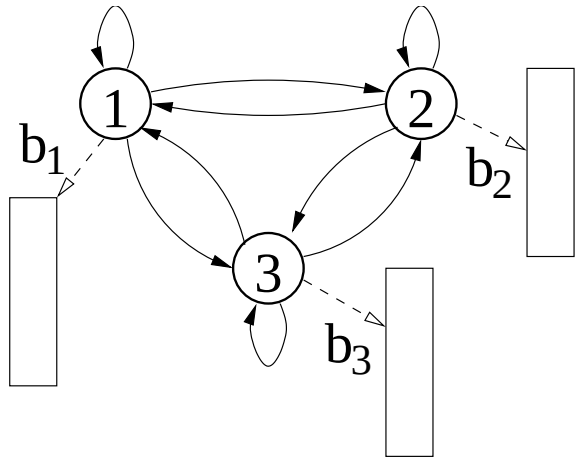
$$A = \left[\begin{array}{c|ccc|c} 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0.6 & 0.3 & 0.1 & 0 \\ 0 & 0 & 0.9 & 0.1 & 0 \\ 0 & 0 & 0 & 0.8 & 0.2 \\ \hline 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

Probability of a given state sequence X:

$$P(X|\mathcal{M}) \;=\; \pi_{x_1} \left(\prod_{t=2}^{T} a_{x_{t-1}x_t} \right) \eta_{x_T} \qquad (17)$$

Introduction to hidden Markov models

Hidden Markov Models (HMMs) use a Markov chain to model stochastic state sequences which emit stochastic observations, e.g., the state topology of an ergodic HMM:



Probability of state i generating **discrete** observation o_t, which has a value from a finite set $k \in 1..K$, is

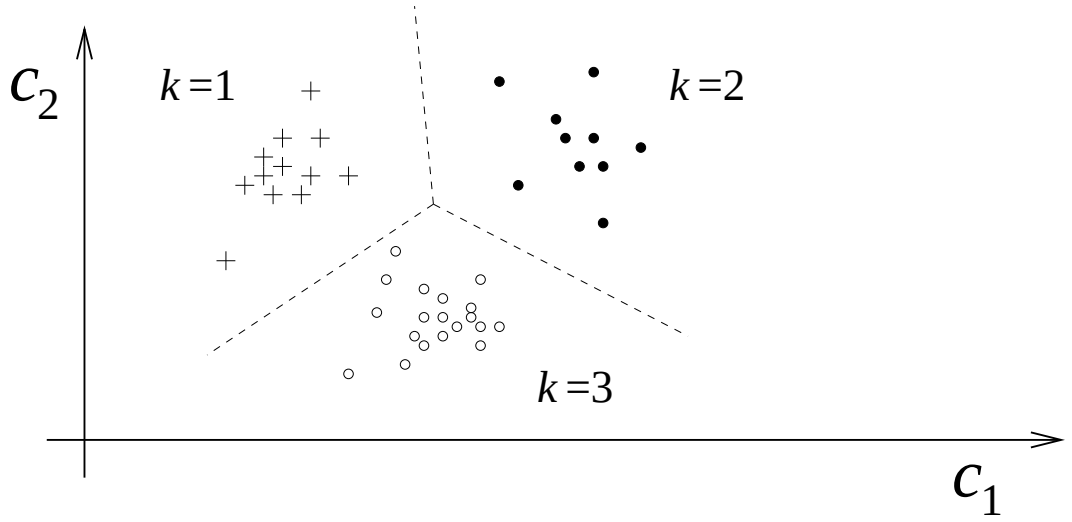
$$b_i(o_t) = P(o_t = k | x_t = i) \tag{18}$$

Probability *distribution* of a **continuous** observation o_t, which takes a value from an infinite set, is

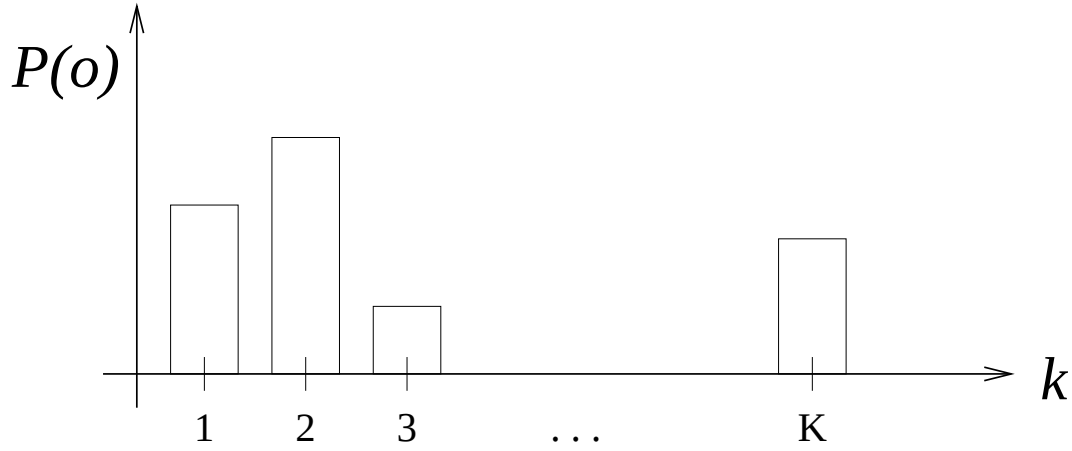
$$b_i(o_t) = p(o_t | x_t = i) \tag{19}$$

We begin by considering only discrete observations.

Observations in discretised feature space



Discrete output probability histogram



Parameters of a discrete HMM, λ

State transition probabilities,

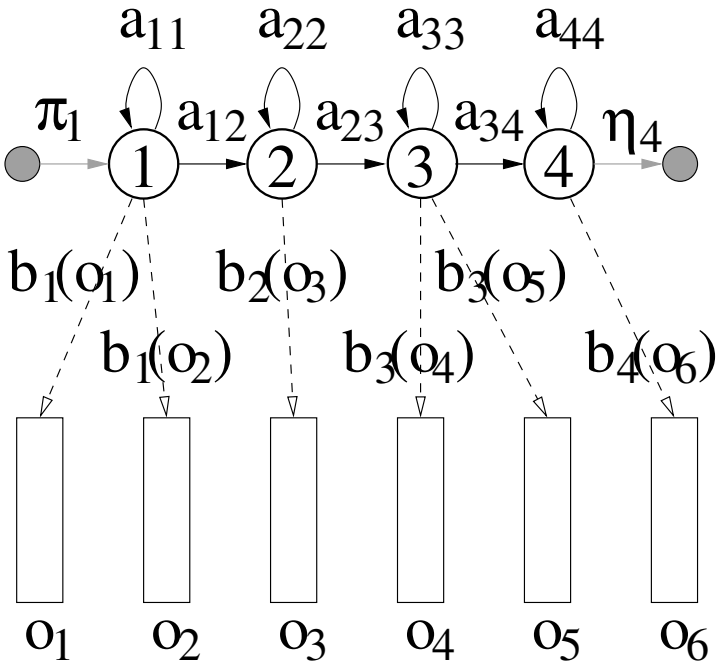
$A = \{\pi_j, a_{ij}, \eta_i\} = \{P(x_t = j | x_{t-1} = i)\}$ for $1 \leq i, j \leq N$

where N is the number of states

Discrete output probabilities,

$B = \{b_i(k)\} = \{P(o_t = k | x_t = i)\}$ for $1 \leq i \leq N$
$1 \leq k \leq K$

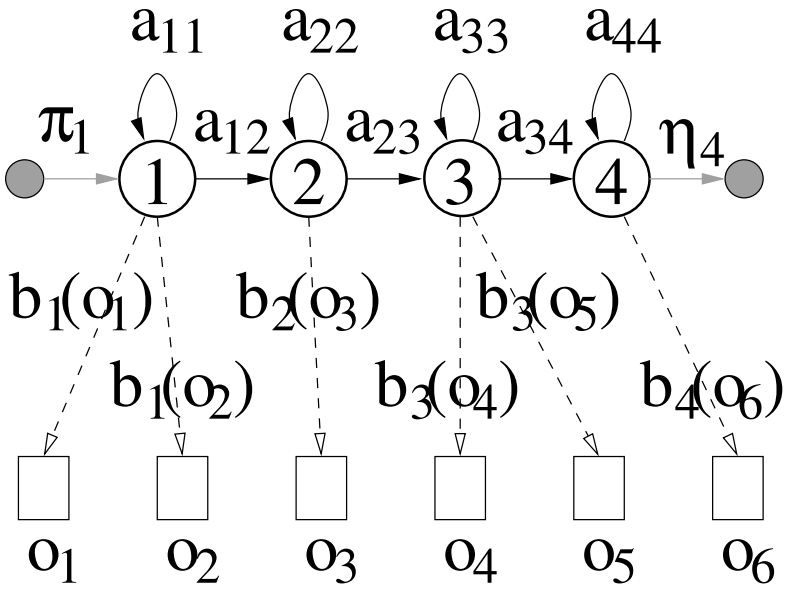
where K is the number of observation types



generating
a state sequence
$X = \{1, 1, 2, 3, 3, 4\}$
and observations
$\mathcal{O} = \{o_1, o_2, \ldots, o_6\}$

Procedure for generating an observation sequence

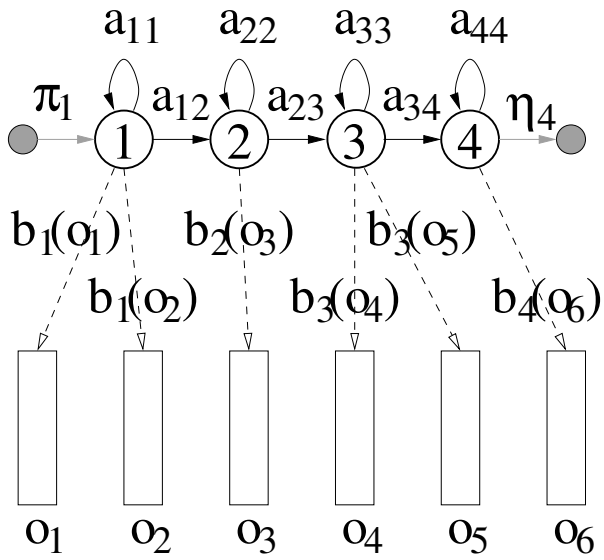
1. For $t = 1$, choose state $x_t = i$ using entry probability π_i
2. Select $o_t = k$ according to $b_{x_t}(k)$
3. Transit according to a_{ij} and η_i, then respectively:
   (a) increment t, set $x_t = j$ and repeat from 2, or
   (b) terminate the sequence, $t = T$.



HMM probability calculation

The joint likelihood of state and observation sequences is

$$P(\mathcal{O}, X|\lambda) = P(X|\lambda)\, P(\mathcal{O}|X, \lambda) \tag{20}$$

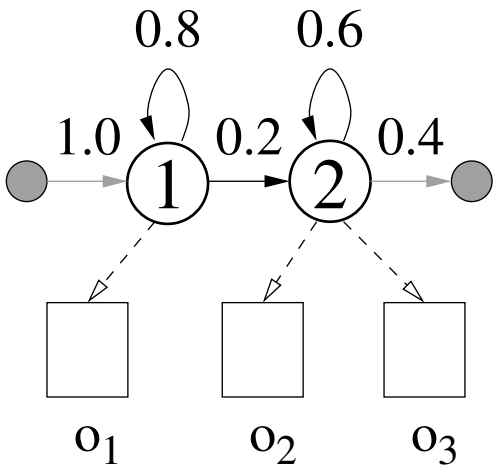


The state sequence $X = \{1, 1, 2, 3, 3, 4\}$ produces the set of observations $\mathcal{O} = \{o_1, o_2, \ldots, o_6\}$:

$$\begin{aligned} P(X|\lambda) &= \pi_1\, a_{11}\, a_{12}\, a_{23}\, a_{33}\, a_{34}\, \eta_4 \\ P(\mathcal{O}|X, \lambda) &= b_1(o_1)\, b_1(o_2)\, b_2(o_3)\, b_3(o_4)\, b_3(o_5)\, b_4(o_6) \\ P(\mathcal{O}, X|\lambda) &= \pi_{x_1} b_{x_1}(o_1) \left(\prod_{t=2}^{T} a_{x_{t-1} x_t} b_{x_t}(o_t) \right) \eta_{x_T} \end{aligned} \tag{21}$$

Example: probability of HMM state sequence

Consider state topology and state transition matrix:



$$A = \left[\begin{array}{c|cc|c} 0 & 1 & 0 & 0 \\ \hline 0 & 0.8 & 0.2 & 0 \\ 0 & 0 & 0.6 & 0.4 \\ \hline 0 & 0 & 0 & 0 \end{array}\right]$$

Output probabilities:

$$B = \left[\begin{array}{c} b_1 \\ b_2 \end{array}\right] = \begin{array}{c} \begin{array}{ccc} \text{R} & \text{G} & \text{B} \end{array} \\ \left[\begin{array}{ccc} 0.5 & 0.2 & 0.3 \\ 0 & 0.9 & 0.1 \end{array}\right] \end{array}$$

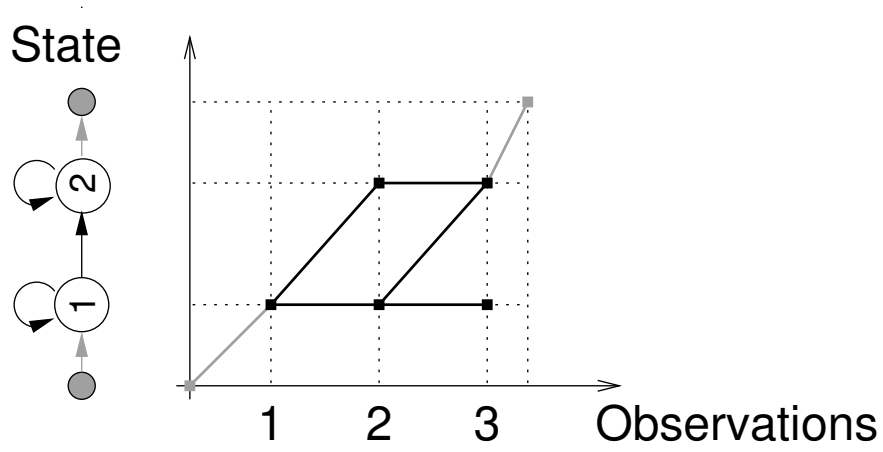
Probability of observations with state sequence $X = \{1, 2, 2\}$:

$$\begin{aligned} P(\mathcal{O}, X|\lambda) &= P(X|\lambda)\, P(\mathcal{O}|X, \lambda) \\ &= \pi_1 b_1(o_1)\, a_{12} b_2(o_2)\, a_{22} b_2(o_3)\, \eta_2 \\ &= 1 \times \\ &= \end{aligned}$$

HMM Recognition & Training

Three tasks within HMM framework

1. Compute likelihood of a set of observations with a given model, $P(\mathcal{O}|\lambda)$
2. Decode a test sequence by calculating the most likely path, X^*
3. Optimise pattern templates by training parameters in the models, $\Lambda = \{\lambda\}$



HMM part 1 summary

- Probability fundamentals
  - Normalisation and marginalisation
  - Joint and conditional probabilities
  - Bayes' theorem
- Markov models
  - sequence of directly observable states
- Hidden Markov models (HMMs)
  - hidden state sequence
  - generation of observations