An introduction to Markov and Hidden Markov Models

Anders Meng

Oktober 2003

1 Introduction

This note is made as an gentle introduction to Markov Models and Hidden Markov Models, and should in combination with ref. [1] give a good overview of the principles of Hidden Markov Models. Inspiration have been found in ref. [4] and ref. [1], both describing the theory of Markov and Hidden Markov Models. To read this note basic statistical skills is an advantage. Especially Bayes formula should be known.

2 Markov Models

Until now in the course we have been assuming that data is independent and identically distributed (iid.). This means that given a sequence of data $\mathbf{D} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 ..., \mathbf{x}_N\}$, sampled from the random variable $\mathcal{X}$, the likelihood can be written as the product of the individual samples

$$p(\mathbf{D}|\mathcal{M}) = \prod_{n=1}^{N} p(\mathbf{x}_n). \tag{1}$$

When working with sequential data (correlation among subsequent samples) the iid. assumption may no longer be a good approximation. Many signals of interests are sequential in nature, eg. audio (speech, music) and video-recordings etc. In both cases it is obvious that there is a time correlation between the different samples.

Instead of assuming independence, we could assume a causal dependence among the given samples.

$$\begin{aligned} p(\mathbf{D}|\mathcal{M}) &= p(\mathbf{x}_N, \mathbf{x}_{N-1}, .., \mathbf{x}_2, \mathbf{x}_1) \\ &= p(\mathbf{x}_N|\mathbf{x}_{N-1}, .., \mathbf{x}_2, \mathbf{x}_1)p(\mathbf{x}_{N-1}|\mathbf{x}_{N-2}.., \mathbf{x}_2, \mathbf{x}_1)....p(\mathbf{x}_2|\mathbf{x}_1)p(\mathbf{x}_1) \quad (2) \\ &= p(\mathbf{x}_1)\prod_{n=2}^{N} p(\mathbf{x}_n|\mathbf{x}_{1:n-1}). \quad (3) \end{aligned}$$

This would be a computational hard problem as one would need a big amount of realizations to find the parameters of this distribution. To simplify the above calculation a *Markov Assumption* is applied

$$p(\mathbf{x}_n|\mathbf{x}_{n-1}, ..., \mathbf{x}_2, \mathbf{x}_1) \approx p(\mathbf{x}_n|\mathbf{x}_{n-1}). \tag{4}$$

This approximation is called the *first order Markov assumption* because the outcome of x_n is only dependent on the outcome at x_{n-1}. The assumption means that eq. (2) can be written as a product of conditions on the previous sample

$$p(\mathbf{D}|\mathcal{M}) = p(\mathbf{x}_1)\prod_{n=2}^{N} p(\mathbf{x}_n|\mathbf{x}_{n-1}). \tag{5}$$

When working with Markov Models the observed variable $\mathcal{X}$ are in some cases[1] discretized using eg. a quantizer, such that the observation sequence $\mathbf{x}_1, \mathbf{x}_2, .., \mathbf{x}_N$ can be described using a scalar discrete observation sequence $\{x_1, x_2, .., x_N\}$ where

[1] The Kalman Filter is considering continuous state space instead of discrete state space.

each of the variables x_n may take one of M states $\{S_1, S_2, .., S_M\}$ (one of M different quantization levels). The likelihood of the discretized samples $D = \{x_1, x_2, .., x_N\}$ can now be calculated as

$$p(D|\mathcal{M}) = p(x_1 = S_i) \prod_{n=2}^{N} p(x_n = S_j | x_{n-1} = S_i). \tag{6}$$

The conditional probabilities $p(x_n = S_j | x_{n-1} = S_i)$ are referred to as *state transition probabilities* or simply *transition probabilities*. The transition probabilities describes the probability of being in state S_j at time $n+1$ given that we where in state S_i at time n :

$$a_{i,j} = p(x_n = S_j | x_{n-1} = S_i). \tag{7}$$

In most cases we assume that the transition probabilities are homogeneous, which means that the probabilities do not change over time, so

$$p(x_n = S_j | x_{n-1} = S_i) = p(x_{n+T} = S_j | x_{n-1+T} = S_i). \tag{8}$$

where T is a positive integer larger or equal to one.

The transition probabilities can be written as a transition matrix, which is of dimension $M \times M$ ($M = 3$ in the example below)

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}. \tag{9}$$

Since each element in the matrix represent a probability of staying or jumping to another state then

- Each entry must be positive, so $a_{i,j} \geq 0$ for all i, j
- Each row must sum up to one, since each row represents the probability of jumping from or staying in the state. Hence $\sum_{j=1}^{M} a_{i,j} = 1$ for $i = 1..M$

The state and transition probabilities can be shown graphically, see figure 1 page 4 showing a three state model, where the transition probabilities is applied.

There is no time information in this illustration. To include the time information another representation is needed, see figure 2 page 4. This way of representation is

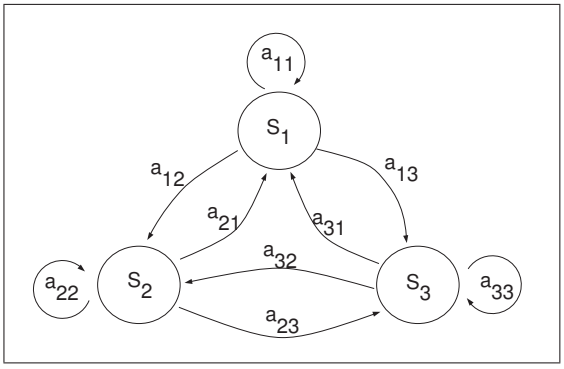

Figure 1: *Illustration of the Markov Model*

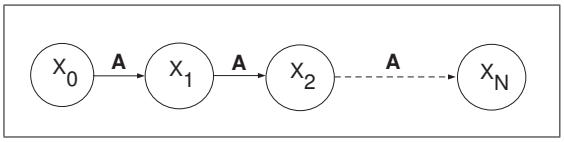

Figure 2: *Graphical model of Markov Model*

called a graphical model. Here each node represents the observed state variable x_n, and the transition matrix is shown between each time-step.

To fully characterize the Markov Model, we need to address the *initial state probability* which is given as $\pi_i(1) = p(x_1 = S_i)$. This may also be written as a vector : $\boldsymbol{\pi}(1) = [p(x_1 = S_1) \;\; p(x_1 = S_2) \;\; . \;\; . \;\; p(x_1 = S_M)]^T$. This probability defines the probability of being in one of the M states at the first sample.

We now have enough knowledge to present the first example : *the weather example.*

2.1 Weather Example

Lets assume that we observe only three different kinds of weather, namely *sunny*, *rainy* or *foggy* weather. We will now use a Markov Model to model the weather. The Markov Model can be build using three states which is given by $\{S_1 = \text{sunny}, S_2 = \text{rainy}, S_3 = \text{foggy}\}$. The stochastic variable $\mathcal{X}$ is a discrete random variable (a scalar) taking one of these three values. A nice old man with big white beard, has provided us with the transition matrix, see table 1 page 5, which can be illustrated as shown in figure 3 page 5.

Having defined the transition matrix, we can now answer questions like : *Given that today is sunny, what's the probability that tomorrow is sunny and the day after is*

		Tomorrow's weather		
		Sunny (S_1)	Rainy (S_2)	Foggy (S_3)
Today's weather	Sunny (S_1)	0.8	0.05	0.15
	Rainy (S_2)	0.2	0.6	0.2
	Foggy (S_3)	0.2	0.3	0.5

Table 1: Transition probabilities

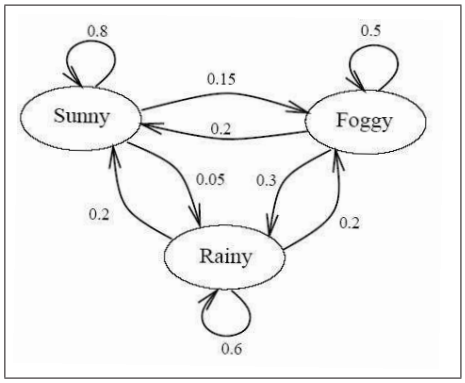

Figure 3: *States and Transition probabilities of the weather Markov Model*

rainy ?

So basically we want to determine the probability : $p(x_2 = Sunny, x_3 = Rainy|x_1 = Sunny) = p(x_2 = S_1, x_3 = S_2|x_1 = S_1)$

$$\begin{aligned} p(x_2 = S_1, x_3 = S_2|x_1 = S_1) &= p(x_3 = S_2|x_2 = S_1, x_1 = S_1)p(x_2 = S_1|x_1 = S_1) \quad (10) \\ &= p(x_3 = S_2|x_2 = S_1)p(x_2 = S_1|x_1 = S_1) \quad (11) \\ &= 0.05 \cdot 0.8 \\ &= 0.04. \end{aligned}$$

The Markov property was applied in eq. (10) to get to eq. (11).

Another question which might be relevant to ask is : *Given that today is foggy, what's the probability that it will be rainy two days from now ?*

Expressed using probabilities this looks like

$$p(x_3 = rainy|x_1 = foggy) = \sum_{x_2} p(x_3 = rainy, x_2|x_1 = foggy).$$

We here marginalize day 2 since there are three ways to get to rainy. The three paths are {foggy,sunny,rainy}, {foggy,rainy,rainy} and {foggy,foggy,rainy}. Using the Markov property the sum can be written as

$$\begin{aligned} p(x_3 = S_2|x_1 = S_3) &= \sum_{x_2=\{S_1,S_2,S_3\}} p(x_3 = S_2|x_2)p(x_2|x_1 = S_3) \\ &= 0.3 \cdot 0.5 + 0.6 \cdot 0.3 + 0.05 \cdot 0.2 \\ &= 0.34. \end{aligned} \tag{12}$$

In the above two examples the start state is given, namely as sunny in the first example and foggy in the second example.

2.2 Finding parameters of the model

As seen from this simple weather example, a nice old man with white beard provided us with the correct transition matrix, but what should we do when we are not provided with the correct matrix ? Our model is described by the parameters $\mathcal{M} = \{\mathbf{A}, \boldsymbol{\pi}\}$, where $\boldsymbol{\pi}$ is the initial state probability. Given we have observed a sequence (training set) $D = \{x_1, x_2, .., x_N\}$ the log-likelihood can be maximized with respect to the initial state probabilities and transition probabilities

$$\arg_{\mathcal{M}} logp(D|\mathcal{M}) = \arg_{\mathcal{M}} \log \left[p(x_1) \prod_{n=1}^{N} p(x_n|x_{n-1}) \right] \tag{13}$$

$$= \arg_{\mathcal{M}} \log[p(x_1)] + \sum_{n=1}^{N} \log[p(x_n|x_{n-1})]. \tag{14}$$

This maximization can be carried out by taking derivatives of the log-likelihood with respect to the transition probabilities as well as the initial state probability. Though the problem of maximizing the log-likelihood is not an unconstrained maximization problem, since the both the transition probability as well as the initial state probabilities have to be positive and sum to one. To overcome this problem Lagrange multipliers or a re-parametrization like softmax (see, eg. [2]) can be used.

2.3 Summary

A small introduction have been given to Markov Models. At this stage, you should know what a Markov Model is. You have been introduced to the following important subjects which characterize a Markov Model :

- identified the difference between independent data and sequential data
- been introduced to a first order Markov Model
- been introduced to states in a Markov Model as well as transition probabilities (transition matrix)
- been introduced to a simple example of a Markov Model where the transition matrix has been given.

2.4 Questions

Describe in words what is meant with sequential data ?

How can you verify that a transition matrix is valid ?

Using the first order Markov assumption would the probability $p(x_n|x_{1:n-1})$ *be influenced by a change in the outcome of* x_{n-3}*?*

3 Hidden Markov Models (HMM)

Now, lets assume that the states $\{S_1, S_2, ..., S_M\}$ which have been observed through the variable $\mathcal{X}$ earlier is no longer observable but are now hidden variables. The hidden variables are observed through another variable $\mathcal{Y}$, which gives an indication of the hidden state. In this discussion we will again concentrate on first order Markov Models. Before continuing we will give an example on how to interpret hidden and observed variables.

> Assume the following scenario. You are observing dice throws at a casino. The person throwing the dice can either use a dice (D_e) with equal probabilities or a dice (D_{ue}) with unequal probabilities. You as the observer, only see the outcome of the different throws without knowledge of which dice was used. In this scenario the outcomes of the different throws are the observed variable and the dice used (either D_e or D_{ue}) is the hidden variable. So an observation sequence could perhaps look like
>
> $$Y = \{453346334454532\}.$$
>
> And the problem could be to determine the most likely sequence of hidden states.

The first order Markov property now applies to the hidden states so, *given the state of $x_{n-1} = S_i$ the current state $x_n = S_j$ is independent of all the states prior to $n-1$.* This means that the transition probabilities (transition matrix) is defined as follows for a first order Hidden Markov Model

$$p(x_n = S_j | x_{n-1} = S_i) = a_{i,j}. \tag{15}$$

The observed variable $\mathcal{Y}$ can be a discrete or continuous random variable. Assume that the observable's are discrete taking on one of L values $\{C_1, C_2, .., C_L\}$, we can relate the observable's and hidden state variables, using the conditional probability:

$$p(y_n = C_i | x_n = S_j) = b_{i,j} \tag{16}$$

which means that given we are in state S_j of the hidden variable at some time instant n, what is the probability of observing the symbol C_i. The output (observables) also satisfy the Markov property with respect to the states: *given x_n the observed output variable y_n is independent of the states and observations at all other time instants.* From eq. (16), it should be easy to convince yourself that the output model

(also called emission probability) can be fully described by a $L \times M$ observation (or emission) matrix $\mathbf{B}$. In most cases we assume that the observation matrix is homogeneous, namely that it do not change over time.

As to recap, the difference between an ordinary Markov Model, and a Hidden Markov Model is that the states is observed directly in the Markov Model, and observed indirectly with a uncertainty in the Hidden Markov Model.

The above is best illustrated using the graphical model representation, see figure 4 page 9 where the hidden variables are hatched[2].

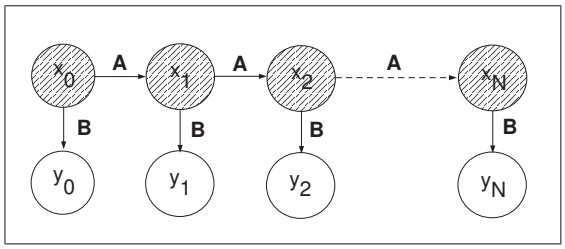

Figure 4: *Graphical model of a Hidden Markov Model*

In the Markov Model we could write out the complete likelihood, since all the variables have been observed. To describe the likelihood of the Hidden Markov Model, we have to remove all uncertainty on the hidden states $\mathcal{X}$, which is done by inference (summing out the underlying states). To calculate the likelihood of the set of observations $D = \{y_1, y_2, .., y_N\}$ using the above rules gives

$$p(D) = \sum_{x_1,x_2,..,x_N} p(x_1)p(y_1|x_1) \prod_{n=2}^{N} p(x_n|x_{n-1})p(y_n|x_n). \tag{17}$$

The above expression can also be written using the notation of transition matrix, observation matrix and initial state probability

$$p(D) = \sum_{x_1,x_2,..,x_N} \pi_{x_1}(1)b_{x_1}(y_1) \prod_{n=2}^{N} a_{x_{n-1},x_n} b_{x_n}(y_n) \tag{18}$$

where $b_{x_n=S_j}(y_n = C_i) = b_{j,i}$ is the emission probability (the probability of emitting symbol C_i given we are in state S_j.

From eq. (18) we see that we are performing N-sums, where each sum is over the number of states M. If one just performed the sums without using the Markov

[2]The graphical model of a Kalman filter is identical with the graphical model of a hidden Markov Model.

	Probability of umbrella given the weather
sunny, $b_{x_n=sunny}(y_n = C_1)$	0.1
rainy, $b_{x_n=rainy}(y_n = C_1)$	0.8
foggy, $b_{x_n=foggy}(y_n = C_1)$	0.3

Table 2: Probability of person bringing an umbrella given the weather

properties, it would require around $\mathcal{O}(M^N)$ operations, which even in the simple case of $M = 10$ and $N = 100$ would require around 10^{100} operations. After a continuation of the weather example we will introduce algorithms to reduce the number of operations to calculate eg. the likelihood. The number of operations can be drastically reduced ($\mathcal{O}(N * M^2)$) due to the first order Markov assumption.

As to explain the Hidden Markov Model, using the weather example from before, we now consider the following modification to the weather problem [3]:

Earlier it was possible for us to observe the weather directly, now assume that you have been locked into a room for several days, and you have no windows, so the only way you can assume something about the weather outside is by observing if the person who brings food to you (hereafter called caretaker) have brought an umbrella. The following probabilities apply (see table 2 page 10)

So the observable variable $\mathcal{Y}$ can only take two values namely $\{C_1 = \text{umbrella}, C_2 = \text{no umbrella}\}$. The hidden variable is now the weather, so $\mathcal{X}$ consist of three states $\{S_1 = \text{sunny}, S_2 = \text{rainy}, S_3 = \text{foggy}\}$. Now we assume that the values of the transition matrix have not been changed, so it is now possible for us to answer questions like,

suppose that the day you where locked into the room it was sunny. The next day the caretaker carried an umbrella into the room, now what is the probability that today is rainy?

This question can be posed using probabilities :

$$p(x_2 = S_2 | y_2 = C_1, x_1 = S_1) \tag{19}$$

And answered by the help of Bayes formula and the Markov property

$$
\begin{aligned}
p(x_2 = S_2 | y_2 = C_1, x_1 = S_1) &= \frac{p(x_2 = S_2, y_2 = C_1, x_1 = S_1)}{p(y_2 = C_1, x_1 = S_1)} && (20)\\
&= \frac{p(x_2 = S_2 | x_1 = S_1) p(y_2 = C_1 | x_2 = S_2) p(x_1 = S_1)}{\sum_{x_2} p(y_2 = C_1, x_2, x_1 = S_1)} && (21)\\
&= \frac{p(x_2 = S_2 | x_1 = S_1) p(y_2 = C_1 | x_2 = S_2) p(x_1 = S_1)}{\sum_{x_2} p(y_2 = C_1 | x_2) p(x_2 | x_1 = S_1) p(x_1 = S_1)} && (22)\\
&= \frac{p(x_2 = S_2 | x_1 = S_1) p(y_2 = C_1 | x_2 = S_2)}{\sum_{x_2} p(y_2 = C_1 | x_2) p(x_2 | x_1 = S_1)} && (23)\\
&= \frac{0.05 \cdot 0.8}{0.1 \cdot 0.8 + 0.8 \cdot 0.05 + 0.3 \cdot 0.15} = 0.243. && (24)
\end{aligned}
$$

The transition from eq. (20) to eq. (21) is due to the Markov property. In the denominator the variable x_2 is added and removed by inference using Bayes rule. So there is approximately 25% chance that the day will be rainy.

Another question, which might be relevant is,

still assuming that the day you where locked in it was sunny. You have observed that the caretaker brought an umbrella on day 2 but did not bring an umbrella on day 3, so what is the probability that day 3 is foggy ?

This would still be possible to answer without to many calculations similar to what we did in the simple Markov case. Now consider this hypothetical question : *Given the day you where locked in it was sunny. The following fifty (50) days you have observed if the caretaker have brought an umbrella or not, then what is the probability that day 51 is sunny ?.*

The question posed in probabilities would be $p(x_5 1 = sunny | y_{2:50}, x_1 = sunny)$. Basically one have to sum out all the states between day 1 and day 51 to answer this question, a really hard task!

To overcome this task in reasonable time, the *forward recursion* also denoted as the forward algorithm is introduced.

3.1 The forward recursion

Due to the first order Markov property it is possible to answer questions as the above recursively using the forward algorithm. The forward variable denoted as $\alpha(y_{1:t}, i)$ is the joint probability (for more details see appendix A page 19)

$$\alpha(y_{1:t}, i) = p(y_1, y_2, .., y_t, x_t = S_i). \tag{25}$$

The forward variable can be recursively determined:

1. Initialization

$$\alpha(y_1, i) = \pi_i(1) b_{S_i}(y_1) \text{ for } 1 \leq i \leq M \tag{26}$$

where M still represents the number of hidden states.

2. Recursion

$$\alpha(y_{1:t+1}, j) = \left[\sum_{i=1}^{M} \alpha(y_{1:t}, i) a_{i,j} \right] b_{S_j}(y_{t+1}) \text{ for } t = 1, 2, .., T-1,\ 1 \leq j \leq M. \tag{27}$$

With the forward algorithm it is now possible to determine the likelihood. The likelihood of the observation sequence $\mathcal{Y} = \{y_1, y_2, .., y_T\}$ can then be calculated by summation over the last state x_T.

$$p(\mathcal{Y}|\mathcal{M}) = \sum_{x_T} p(y_{1:T}, x_T) = \sum_{i=1}^{M} \alpha(y_{1:T}, i). \tag{28}$$

The computation of the forward variable can consist of a large number of values that are less than 1. Hence after a few observations the values of the forward variable will head exponentially towards zero, exceeding the floating point precision. The numerical problems introduced can be solved by scaling which will be discussed in section 3.5 page 17.

With the alpha recursion at hand, it is possible in a few lines of code to recursively calculate the question posed earlier. We are interested in calculating the following probability (where $t = 51$):

$$p(x_t = sunny|y_1, y_2, ..., y_t) = \frac{\alpha(y_{1:t}, 1)}{\sum_{i=1}^{M} \alpha(y_{1:t}, i)} \tag{29}$$

The calculation in eq. (29) can be obtained using Bayes rule and the definition of $\alpha(y_{1:t}, i)$.

The problem of answering the question $p(x_t = sunny|y_1, y_2, ..., y_t)$ is also known as filtering in the litterature. The reason for this terminology comes from the interpretation that the observables (or outputs) y_t is providing "noisy" information about the underlying "signal" x_t. The inference problem is then "filtering" the noise from the signal.

The complexity of this algorithm can be determined to be $\mathcal{O}(TM^2)$ assuming we have T observables.

Another question which might be relevant to ask is the following *Given the day you where locked in it was sunny (so $p(x_1 = sunny) = 1$). The following fifty (50) days you have observed if the caretaker was bringing an umbrella or not, then what is the probability that day 20 was sunny ?*

Formulating this question as probabilities we want to answer the following question ($T = 51$ and $t = 20$)

$$p(x_t = sunny|y_1, y_2, .., y_t, .., y_T) \tag{30}$$

One could think, why not just sacrifice the additional samples we have observed and then just run the forward recursion as to determine $p(x_t = sunny|y_{1:t})$ disregarding the additional information ?

The simple answer to this question is that the additional data, which we are given, strengthens our assumptions about the true outcome at that specific day. The problem of determining questions like $p(x_t = sunny|y_{1:T})$ is in the litterature also known as the smoothing (interpolation) problem.

3.2 The backward recursion

To be able to solve the smoothing problem we need to be able to calculate the following probability $p(y_{t+1}, y_{t+2}, .., y_T|x_t = S_i)$. Why this probability you might ask! The simple explanation is given by rewriting the smoothing problem into something recognizable :

$$\begin{aligned}
p(x_t = S_i|y_{1:T}) &= \frac{p(y_{1:T}|x_t = S_i)p(x_t = S_i)}{p(y_{1:T})} && (31)\\
&= \frac{p(y_{1:t}|x_t)p(y_{t+1:T}|x_t = S_i)p(x_t = S_i)}{p(y_{1:T})} && (32)\\
&= \frac{p(y_{1:t}, x_t = S_i)p(y_{t+1:T}|x_t = S_i)}{p(y_{1:T})} && (33)\\
&= \frac{\alpha(y_{1:t}|i)\beta(y_{t+1:T}, i)}{p(y_{1:T})} && (34)
\end{aligned}$$

where we define $\beta(y_{t+1:T}, i) = p(y_{t+1}, y_{t+2}, .., y_T|x_t = S_i)$. In eq. (31) we simply use Bayes rule as to rewrite the smoothing problem in to something dependent on the

underlying state at time t. Due to the first order Markov property the step from eq. (31) to eq. (32) is feasible. The step from eq. (32) to eq. (33) is to generate the joint distribution such that we can write up the expression in eq. (34). We recognize the forward variable $\alpha(y_{1:t}|i)$, which we know how to calculate. The likelihood $p(y_{1:T})$ can be calculated after one complete run with the forward algorithm. The only thing we have to determine to be able to solve the smoothing problem is $\beta(y_{t+1:T}, i) = p(y_{t+1:T}|x_t = S_i)$.

As with the forward algorithm it is possible with the use of Bayes formula and the Markov properties to derive a recursive formula for determining the backward variable. More details can be found in appendix A page 19.

1. Initialization

$$\beta(y_T, i) = 1 \ , \ 1 \leq i \leq M \tag{35}$$

2. Induction

$$\begin{aligned} \beta(y_{t:T}, i) &= \sum_{j=1}^{M} a_{i,j} b_{S_j}(y_{t+1}) \beta(y_{t+1:T}, j) \\ &\quad \text{for } t = T-1, T-2, .., 1 \text{ and } 1 \leq i \leq M \end{aligned} \tag{36}$$

The computational requirements of the backward recursion is similar to the forward recursion, $\mathcal{O}(T \times M^2)$ operations.

As with the forward algorithm there will be problems with the floating point precision when calculating the backward variable. As to overcome these problems scaling is needed. This issue will be discussed in section 3.5 page 17.

3.3 Estimation of parameters in a HMM

In both the weather example and the coming coin-tossing example we are given the initial state probabilities, transition probabilities and emission probabilities. In the general case we do not know these probabilities, so we have to use a training set to determine these parameters.

As with the Markov-Model estimation of parameters for the HMM can be done by maximizing the likelihood function. Due to the fact that we are working on incomplete dataset since some of the variables are hidden (the hidden states) we can

use the EM-algorithm to find the parameters of the model. Using the EM-algorithm one finds update formulas for the initial state probabilities, transition probabilities as well as the emission probabilities. In ref. [1] update formulas is given for both discrete and continuous emission probabilities.

3.4 Tossing coin example

The coin tossing example (see, eg. [1]) is a good example to illustrate what you have just learned. The plots as well as the code for the plots have been generated using *MATLAB-code* provided in the course.

You are placing bets on the outcome of coin-throws. The person throwing the coins, shifts between two coins with a certain probability: The probability of shifting from coin 1 to coin 2 is 0.05 and the probability of changing from coin 2 to coin 1 is 0.1. The difference between the coins is that one is biased (coin 2, $p(head) = 0.9$) and the other is unbiased (coin 1,$p(head) = 0.5$).

We do not have any prior information on which coin is thrown the first time, so the initial state probability must be $p(x_1 = coin2) = 0.5$. The scenario can be sketched, see figure 5 page 15.

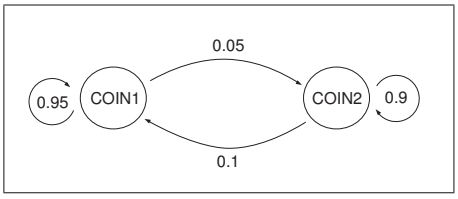

Figure 5: *States and Transition probabilities of throwing coins example*

From the above information the transition matrix, emission matrix and the initial state probability vector can be determined

$$\mathbf{A} = \begin{bmatrix} 0.95 & 0.05 \\ 0.1 & 0.9 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 0.5 & 0.9 \\ 0.5 & 0.1 \end{bmatrix} \quad \boldsymbol{\pi}_1 = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}.$$

We now observe 200 outcomes of the experiment, where the coins have been thrown according to the above probabilities.

One outcome may look like figure 6 page 16, where both the observed variable and hidden variables have been shown (normally we do not know the true underlying states). The person throwing the coins will not tell you when he is using the biased coin, however applying what we know on Hidden Markov Models, the posterior

probability $p(x_t = coin2|y_{1:T})$ for $t = 1..T$ provides us the probability that coin 2 was used, given the observed sequence $y_{1:T}$.

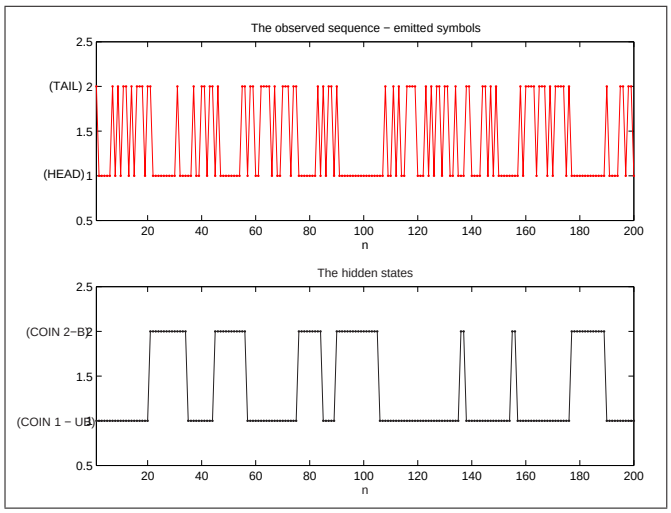

Figure 6: *A simulation run using the above model assumptions*

In figure 7 page 16 we plot the probability of coin 2 given the observed sequence.

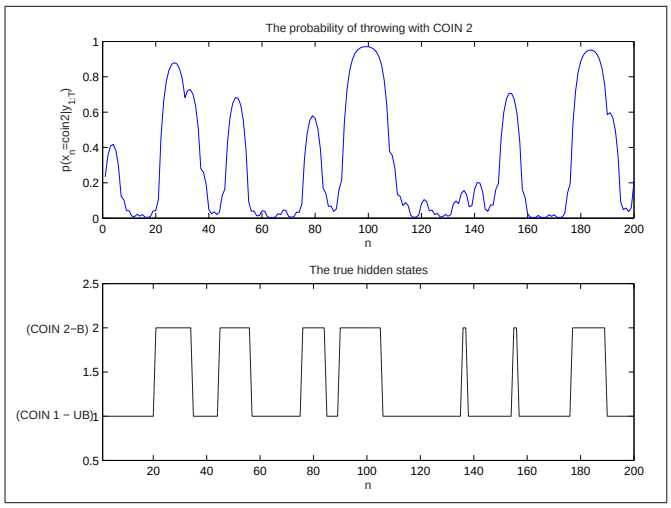

Figure 7: *The posterior probability determined from a forward-backward run*

From the plot we see a very good correlation with the true underlying states, which we do not immediately observe through the observed sequence from upper figure of figure 6 page 16.

The next thing we will investigate is the efficiency of estimating the parameters of

the HMM. This is done using the EM-algorithm. Generating a sequence of throws of length 200 and running the EM-algorithm using a stop-criterium of $log(p(y_{1:T}|\mathcal{M}_k)) - log(p(y_{1:T}|\mathcal{M}_{k-1}) < 1e-5)$ where k is the iteration index on the HMM parameters, gives the following estimates of $\mathbf{A}, \mathbf{B}$ and $\boldsymbol{\pi}_1$

$$\tilde{\mathbf{A}} = \begin{bmatrix} 0.94 & 0.06 \\ 0.07 & 0.93 \end{bmatrix} \quad \tilde{\mathbf{B}} = \begin{bmatrix} 0.48 & 0.95 \\ 0.52 & 0.05 \end{bmatrix} \quad \tilde{\boldsymbol{\pi}_1} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

which is not that far from the true parameters. As to improve the guess a longer observation sequence is needed. The initial state probability is not so close to the true values. The initial state probability though, is only critical during the first states, since the HMM tends to forget what happened N samples ago (the correlation is exponentially decaying), due to the first order Markov property.

3.5 Methods to overcome numerical problems

Until now we have not discussed any practical problems, but there is a problem when implementing the HMM in a computer due to finite precission. In the small examples given, only small alphabets have been used (in the weather example only 3 states was used, and the emission probability could only attain one of two values). When multiplying together a lot of probabilities (numbers between 0 and 1), we might get into problems with machine inaccuracy. As to avoid numerical problems one have to scale the forward variable as well as the backward variable during recursion. The idea of scaling was originally proposed by ref. [1], and basically all you have to do is to change the way you calculate the forward and backward variables. Technical details can be found in ref. [1].

3.6 Summary

You have been introduced to the Hidden Markov Model. This model is more versatile than the normal Markov Model. The Hidden Markov Models are used in many situations where sequential data needs to be modelled. Some examples where Hidden Markov Models is used is bioinformatics, speech-recognition, hand-writing recognition. You have been introduced to the following important subjects

- Observed the difference between a Markov Model and the Hidden Markov Model
- Been introduced to the emission probability

- Recognized that the forward, backward- recursions is fast implementations of the full inference problem
- Found that we need an optimization scheme such as the EM-algorithm to find the initial state probability, transition matrix and emission probability parameters

4 Acknowledgements

The weather story have been presented in quite many papers on Markov models and Hidden Markov Models. It might have been originally proposed in ref. [1], but in ref. [3] it has been refreshed with a HMM example, which is quite informative. Thanks to all the reviewers of this note.

References

[1] L. Rabiner: *A tutorial on Hidden Markov Models and selected applications in speech recognition*, Proceedings of the IEEE 77(2):257-286, 1989.

[2] C. M. Bishop : *Neural Networks for Pattern Recognition*, Oxford University Press 1995. (ISBN : 0 19 853864 2 Paperback)

[3] E. Fosler-Lussier : *Markov Models and Hidden Markov Models : A Brief Tutorial*, Technical Report (TR-98-041), December 1998, International Computer Science Institute, Berkeley, California.

[4] R. Durbin, S. Eddy, A. Krogh and G. Mitchison: *Biological Sequence Analysis - Probalistic Models of proteins and Nuceleic Acids*, Cambridge University Press 1998. (ISBN : 0 521 62041 4 Hardback)

A The forward, forward backward algorithm in probabilities

In this appendix we will derive define the forward - variable $\alpha(y_{1:t}, i)$ and the backward variable $\beta(y_{t+1:T}, i)$ for $i = 1, 2, 3, .., M$. In many cases we are interested in calculating the conditional probability $p(x_t|y_{1:T})$. The conditional probability was also calculated in section 3.2 page 13 but are repeated here to get an overall impression. Using Bayes rule, we can rewrite this probability

$$
\begin{aligned}
p(x_t|y_{1:T}) &= \frac{p(y_{1:T}|x_t)p(x_t)}{p(y_{1:T})} && (37)\\
&= \frac{p(y_{1:t}|x_t)p(y_{t+1:T}|x_t)p(x_t)}{p(y_{1:T})} && (38)\\
&= \frac{p(y_{1:t}, x_t)p(y_{t+1:T}|x_t)}{p(y_{1:T})} && (39)\\
&= \frac{\alpha(x_t)\beta(x_t)}{p(y_{1:T})} && (40)
\end{aligned}
$$

where $\alpha(x_t) = p(y_0, y_1, .., y_t, x_t)$ is the forward variable and expresses the probability of emitting the partial sequence of outputs $y_0, y_1, .., y_t$ and ending up in state x_t. The backward variable $\beta(x_t) = p(y_{t+1}, .., y_T|x_t)$ is the probability of emitting the partial sequence of outputs $y_{t+1}, .., y_T$ given that the systems starts in state x_t. Using the conditional independence in Markov Models it is possible to split the probability in eq. (37) to eq. (38). In eq. (39) the terms have been regrouped by multiplication of $p(x_t)$.

It is now possible with the definition of the forward variable to derive the recursive formulas using the conditional independence property[3] as well as Bayes rule

[3] Conditioning on a state

$$
\begin{aligned}
\alpha(x_{t+1}) &= p(y_0, .., y_{t+1}, x_{t+1}) && (41)\\
&= p(y_0, .., y_{t+1}|x_{t+1})p(x_{t+1}) && (42)\\
&= p(y_0, .., y_t|x_{t+1})p(y_{t+1}|x_{t+1})p(x_{t+1}) && (43)\\
&= p(y_0, .., y_t, x_{t+1})p(y_{t+1}|x_{t+1}) && (44)\\
&= \sum_{x_t} p(y_0, .., y_t, x_t, x_{t+1})p(y_{t+1}|x_{t+1}) && (45)\\
&= \sum_{x_t} p(y_0, .., y_t, x_{t+1}|x_t)p(x_t)p(y_{t+1}|x_{t+1}) && (46)\\
&= \sum_{x_t} p(y_0, .., y_t|x_t)p(x_{t+1}|x_t)p(x_t)p(y_{t+1}|x_{t+1}) && (47)\\
&= \sum_{x_t} p(y_0, .., y_t, x_t)p(x_{t+1}|x_t)p(y_{t+1}|x_{t+1}) && (48)\\
&= \sum_{x_t} \alpha(x_t)p(x_{t+1}|x_t)p(y_{t+1}|x_{t+1}) && (49)\\
&= \sum_{x_t} \alpha(x_t)a_{x_{t+1}|x_t}b_{x_t+1}(y_{t+1}). && (50)
\end{aligned}
$$

The last expression, eq. (50), is the same update formula as given in eq. (27). In eq. (47) remember that $p(y_0, .., y_t|x_{t+1}, x_t) = p(y_0, .., y_t|x_t)$.

It is also possible to determine the recursive update for the backward variable, again using the conditional independence and Bayes rule

$$
\begin{aligned}
\beta(x_t) &= p(y_{t+1}, .., y_T|x_t) && (51)\\
&= \sum_{x_{t+1}} p(y_{t+1}, .., y_T, x_{t+1}|x_t) && (52)\\
&= \sum_{x_{t+1}} p(y_{t+1}, .., y_T|x_{t+1}, x_t)p(x_{t+1}|x_t) && (53)\\
&= \sum_{x_{t+1}} p(y_{t+2}, .., y_T|x_{t+1})p(y_{t+1}|x_{t+1})p(x_{t+1}|x_t) && (54)\\
&= \sum_{x_{t+1}} \beta(x_{t+1})b_{x_{t+1}}(y_{t+1})a_{x_{t+1},x_t}. && (55)
\end{aligned}
$$

In eq. (54) remember that $p(y_{t+2}, .., y_T|x_{t+1}, x_t) = p(y_{t+2}, .., y_T|x_{t+1})$. The expression obtained in eq. (55) is the same expression as for the backward recursion given in eq. (36).