6.867: Recitation Handout (Week 11)

November 22, 2016

Contents

1 DNA **2**

2 Two-state HMM **4**

3 Hidden Where? **5**

4 Two by Two **7**

5 Missing Observations **8**

6 Studiousness **9**

7 IO HMM **11**

8 Partially Observed Markov Chain **12**

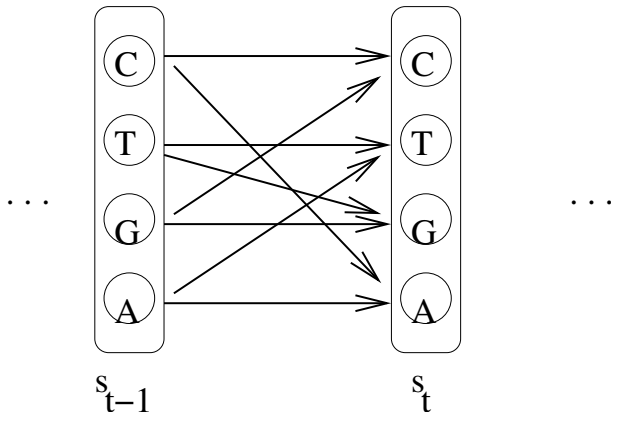
1 DNA

Consider the following pair of observed sequences:

```
Sequence 1 (s_t): AATTGGCC AATTGGCC ...
Sequence 2 (x_t): 11221122 11221122 ...
Position t:       01234 ...
```

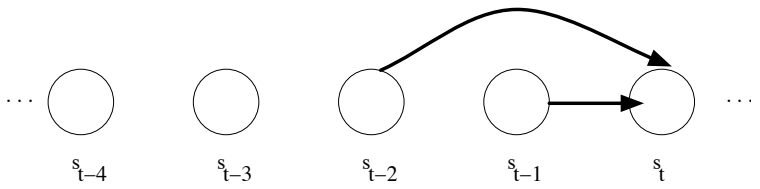
where we assume that the pattern (indicated with the spaces) will continue forever. Let $s_t \in \{A, G, T, C\}, t = 0, 1, 2, \ldots$ denote the variables associated with the first sequence, and $x_t \in \{1, 2\}, t = 0, 1, 2, \ldots$ the variables characterizing the second sequence. So, for example, given the sequences above, the observed values for these variables are $s_0 = A, s_1 = A, s_2 = T, \ldots$ and, similarly, $x_0 = 1, x_1 = 1, x_2 = 2, \ldots$

(a) If we use a simple first-order homogeneous Markov model to predict the first sequence (values for s_t only), what is the maximum likelihood solution that we would find? In the transition diagram below, please draw the relevant transitions and the associated probabilities (this should not require much calculation)

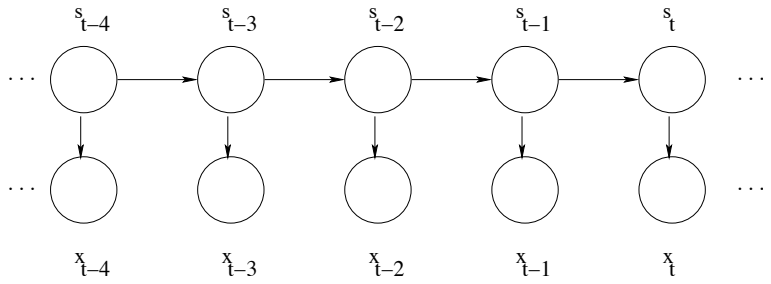


Solution: All the transition probabilities are 0.5.

(b) To improve the Markov model a bit, we would like to define a graphical model that predicts the value of s_t on the basis of the previous observed values $s_{t-1}, s_{t-2}, \ldots$ (looking as far back as needed). The model parameters/structure are assumed to remain the same if we shift the model one step. In other words, it is the same graphical model that predicts s_t on the basis of $s_{t-1}, s_{t-2}, \ldots$ as the model that predicts s_{t-1} on the basis of $s_{t-2}, s_{t-3}, \ldots$. In the graph below, draw the minimum number of arrows that are needed to predict the first observed sequence (s_t) perfectly (disregarding the first few symbols in the sequence). Since we slide the model along the sequence, you can draw the arrows only for s_t.



(c) Now, to incorporate the second observation sequence, we will use a standard hidden Markov model:



where again $s_t \in \{A, G, T, C\}$ and $x_t \in \{1, 2\}$. We will estimate the parameters of this HMM in two different ways.

(A) Treat the pair of observed sequences (s_t, x_t) (given above) as complete observations of the variables in the model and estimate the parameters in the maximum likelihood sense. The initial state distribution $P_0(s_0)$ is set according to the overall frequency of symbols in the first observed sequence (uniform).

(B) Use only the second observed sequence (x_t) in estimating the parameters, again in the maximum likelihood sense. The initial state distribution is again uniform across the four symbols.

We assume that both estimation processes will be successful relative to their criteria.

i. What are the observation probabilities $\Pr(x \mid s)(x \in \{1, 2\}, s \in \{A, G, T, C\})$ resulting from the first estimation approach? (should not require much calculation)

> **Solution:** $\Pr(x = 1 \mid s = A) = 1, \Pr(x = 1 \mid s = G) = 1, \Pr(x = 2 \mid s = T) = 1, \Pr(x = 2 \mid s = C) = 1$, all other probabilities are zero.

ii. Which estimation approach is likely to yield a more accurate model over the second observed sequence (x_t)? Briefly explain why.

> **Solution:** The second one (B); the x_t sequence can be represented exactly by a first-order Markov chain with four states. So, we can use the available four states (A,T,G,C) to exactly capture the variability in the x_t sequence. However, since we're uncertain as to the start state, we'd get a model which assigns probability 1/4 to each observed

sequence of the above type. On the other hand, if we use approach (A), we cannot capture the second-order dependence in the s_t sequence in a standard first-order HMM. Using the observation probabilities above, we'd assign probability 0.5^n for a sequence of length n.

(d) Consider now the two HMMs resulting from using each of the estimation approaches (approaches A and B above). These HMMs are estimated on the basis of the pair of observed sequences given above. We'd like to evaluate the probability that these two models assign to a new (different) observation sequence `1 2 1 2`, i.e., $x_0 = 1, x_1 = 2, x_2 = 1, x_3 = 2$. For the first model, for which we have some idea about what the s_t variables will capture, we also want to know the the associated most likely hidden state sequence. (these should not require much calculation)

i. What is the probability that the first model (approach A) assigns to this new sequence of observations?

Solution: 1/16

ii. What is the probability that the second model (approach B) gives to the new sequence of observations?

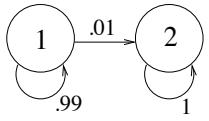
Solution: Zero; the model generates only repeated symbol sequences of the type `1 1 2 2 ...`.

iii. What is the most likely hidden state sequence in the first model (from approach A) associated with the new observed sequence?

Solution: A T G C

2 Two-state HMM

s	1	2
P(x=1)	0	0.1
P(x=2)	0.199	0
P(x=3)	0.8	0.7
P(x=4)	0.001	0.2



The figure above shows a two-state HMM. The transition probabilities of the Markov chain are given in the transition diagram. The output distribution corresponding to each state is defined over 1, 2, 3, 4 and is given in the table next to the diagram. The HMM is equally likely to start from either of the two states.

(a) Give an example of an output sequence of length 2 which can not be generated by this HMM.

Solution: 1,2

(b) We generated a sequence of $6,867^{2016}$ observations from the HMM, and found that the last observation in the sequence was 3. What is the most likely hidden state corresponding to that last observation?

Solution: 2

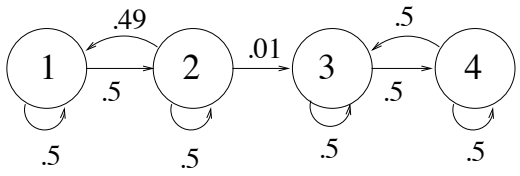
(c) Consider an output sequence 3 3. What is the most likely sequence of hidden states corresponding to these observations?

Solution: 1,1

(d) Now, consider an output sequence 3 3 4. What are the *first two states* of the most likely hidden state sequence?

Solution: 2,2

(e) We can try to increase the modeling capacity of the HMM a bit by breaking each state into two states. Following this idea, we created the diagram below. Can we set the initial state distribution and the output distributions so that this 4-state model, with the transition probabilities indicated in the diagram, would be equivalent to the original 2-state model? If yes, how? If no, why not?



Solution: No we cannot. First note that we have to associate the first two states in the 4-state model with state 1 of the 2-state model. The probability of leaving the first two states in the 4-state model, however, depends on time (whether the chain happens to be in state 1 or 2). In contrast, in the 2-state model the probability of transitioning to 2 is always 0.01.

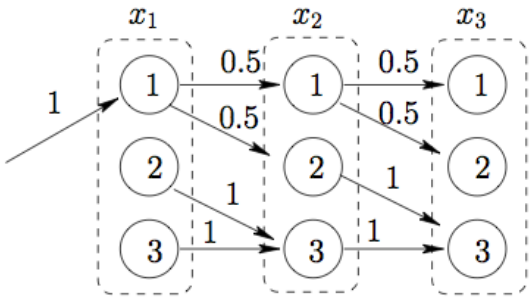
3 Hidden Where?

Suppose we have a hidden Markov model with three possible states and transitions described in the figure below. The states at times $t = 1, 2, 3$ are represented by variables x_1, x_2, and x_3, respectively. The distributions governing binary (0/1) outputs y_1, y_2, and y_3 are state dependent

but do not depend on time. In other words, we only need to define $P(y|x)$ common to all time points. Specifically,

$P(y|x)$:

	$x=1$	$x=2$	$x=3$
$y=0$	0.5	0.5	0.1
$y=1$	0.5	0.5	0.9



(a) There are only three possible hidden state sequences over the three time points $t = 1, 2, 3$. What are they?

> **Solution:** 111, 112, and 123

(b) What is the most likely hidden state sequence given the observed sequence $(y_1 = 0, y_2 = 0, y_3 = 0)$? Is the answer unique?

> **Solution:** Note first that the observation sequence does not help distinguish states 1 and 2 but would discount state 3 since $P(y = 0|x = 3) = 0.1$. The answer is not unique. Both 111 and 112 are equality likely according to the Markov chain and the observations do not help distinguish them. More precisely:
>
> $$\begin{aligned} P(x_1 = 1, x_2 = 1, x_1 = 1, y_1 = 0, y_2 = 0, y_3 = 0) &= (1 \cdot 0.5) \cdot (0.5 \cdot 0.5) \cdot (0.5 \cdot 0.5) \\ P(x_1 = 1, x_2 = 1, x_3 = 2, y_1 = 0, y_2 = 0, y_3 = 0) &= (1 \cdot 0.5) \cdot (0.5 \cdot 0.5) \cdot (0.5 \cdot 0.5) \\ P(x_1 = 1, x_2 = 2, x_3 = 3, y_1 = 0, y_2 = 0, y_3 = 0) &= (1 \cdot 0.5) \cdot (0.5 \cdot 0.5) \cdot (1 \cdot 0.1) \end{aligned}$$
>
> where the probabilities within parenthesis represent transitioning to the state at the corresponding time point and generating $y = 0$.

(c) Suppose we only receive observations for the first two time points, i.e., (y_1, y_2), and train the HMM with the EM algorithm. The HMM is initialized with the parameters illustrated above. Select which of the following parameters could change as a result of running the EM algorithm:

( ) transitions from state $x = 1$, i.e., $P_{1j} = P(x_{t+1} = j|x_t = 1), j = 1, 2, 3$
( ) transitions from state $x = 3$, i.e., $P_{3j} = P(x_{t+1} = j|x_t = 3), j = 1, 2, 3$

( ) output distribution from state $x = 1$, i.e., $P(y|x = 1), y = 0, 1$
( ) output distribution from state $x = 3$, i.e., $P(y|x = 3), y = 0, 1$

Solution: option 1 and 3.

4 Two by Two

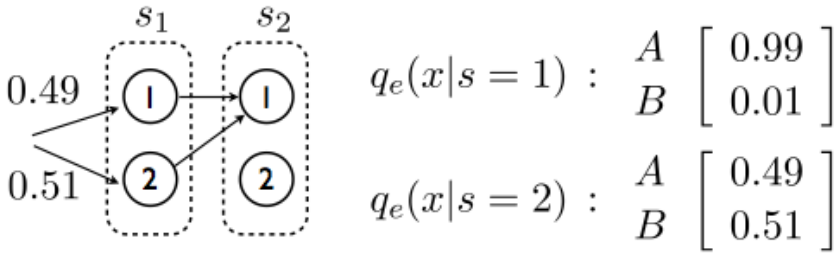


Figure 1: hidden Markov model

Consider a homogeneous hidden Markov model illustrated in Figure 2. There are two values for the hidden state, $s = 1$ and $s = 2$, and two possible output symbols, $x = A$ and $x = B$.

(a) What is the most frequent output symbol (A or B) to appear in the first position of sequences generated from this HMM?

Solution: The most frequent symbol is one that has the highest $P(x_1;)$.

$$\begin{aligned} P(x_1 = A;\theta) &= 0.49 \cdot 0.99 + 0.51 \cdot 0.49 \\ P(x_1 = B;\theta) &= 0.49 \cdot 0.01 + 0.51 \cdot 0.51 \end{aligned}$$

Clearly, $P(x_1 = A;\theta) > P(x_1 = B;\theta)$.

(b) What is the sequence of three output symbols that has the highest probability of being generated from this HMM model?

Solution: AAA

(c) Suppose we generate output sequences of length three from this HMM and also reveal the first state s1. So, each example that we sample is a value assignment to variables s_1, x_1, x_2, and x_3. The special structure in the above HMM permits us to write the following two conditional probabilities as products of $q_e(x|s)$. Provide the expressions.

Solution:

$$\begin{aligned} P(x_1, x_2, x_3|s_1 = 1) &= q_e(x_1|s = 1)q_e(x_2|s = 1)q_e(x_3|s = 1) \\ P(x_1, x_2, x_3|s_1 = 2) &= q_e(x_1|s = 2)q_e(x_2|s = 1)q_e(x_3|s = 1) \end{aligned}$$

(note that the distributions differ only in terms of how x_1 is generated. Thus x_2 and x_3 are independent of s_1)

5 Missing Observations

Say we have a standard HMM, which generates sequences of length $m = 4$, that is the model takes the form

$$p(x_1, x_2, x_3, x_4, s_1, s_2, s_3, s_4) = t(s_1) \prod_{j=2}^{4} t(s_j|s_{j-1}) \prod_{j=1}^{4} e(x_j|s_j)$$

where $t(s_1)$ is the initial probability distribution over the states, $t(s_j|s_{j-1})$ are the state transition probabilities and $e(x_j|s_j)$ are the observation probabilities. Assume now that we condition on $X_2 = 1$ and $X_4 = 1$. We can then consider the conditional distribution

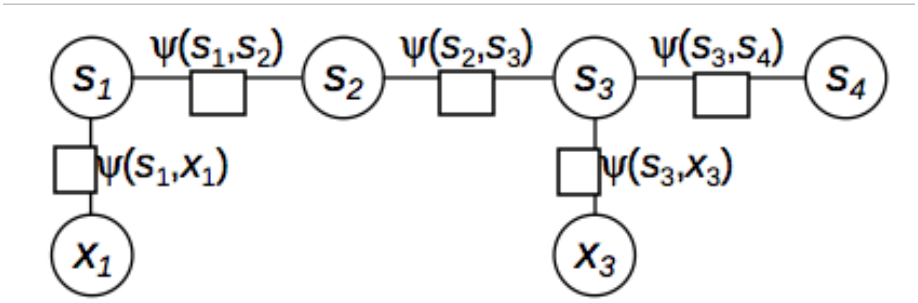
$$p(x_1, x_2, x_3, x_4, s_1, s_2, s_3, s_4|X_2 = 1, X_4 = 1)$$

(a) Define a factor graph over the random variables $X_1, X_3, S_1, S_2, S_3, S_4$ that correctly models the conditional distribution

$$p(x_1, x_2, x_3, x_4, s_1, s_2, s_3, s_4|X_2 = 1, X_4 = 1)$$

You should: (1) show the graph underlying the Markov random field; (2) specify any potential functions ψ on edges in the graph, as functions of the parameters $t(s_1)$, $t(s_i|s_{i-1})$ and $e(x_i|s_i)$.

Solution:



$$\begin{aligned}
\psi(s_1, s_2) &= t(s_1)t(s_2|s_1)e(1|s_2) \\
\psi(s_2, s_3) &= t(s_3|s_2) \\
\psi(s_3, s_4) &= t(s_4|s_3)e(1|s_4) \\
\psi(s_1, x_1) &= e(x_1|s_1) \\
\psi(s_3, x_3) &= e(x_3|s_3)
\end{aligned}$$

A few of the factors can equivalently be moved around, but we need $e(1|s_2)$ and $e(1|s_4)$ in the corresponding factors

(b) Assume that we have an HMM of length m:

$$p(x_1, x_2, ..., x_m, s_1, s_2, ..., s_m) = t(s_1) \prod_{j=2}^{m} t(s_j|s_{j-1}) \prod_{j=1}^{m} e(x_j|s_j)$$

Define a dynamic programming algorithm, that given an input sequence $x_1...x_m$, allows us to efficiently calculate

$$\max_{s_1, s_2, ..., s_m : s_j = s} p(x_1, x_2, ..., x_m, s_1, s_2, ..., s_m)$$

for any $j \in 1...m$, and for any $s \in S$ where S is the set of possible states. That is, the algorithm calculates the maximum probability for any sequence of states under the constraint that s_j is equal to s.

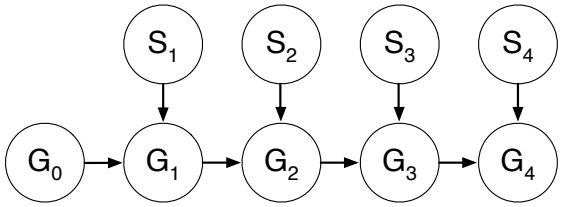
Solution: There are several valid approaches:
1. Using belief propagation: define potentials $\psi_{ab} = t(s_b|s_a)e(x_b|s_b)$, pass messages to position j as the root, combine messages along the way using max. Report $m_{(j-1)\to j}(s)m_{(j+1)\to j}(s)$ (do not divide by partition function). Using terminology typically associated with HMMs, this consists of modifying the forward and backward recurrences to use max instead of sum. Then report $\alpha[j, s]\beta[j, s]$ (do not divide by anything).
2. The commonly used Viterbi algorithm can also be adapted. In this case, we enforce the constraint $s_j = s$ during its pass through the sequence, by forcing any transitions at position j violating the constraint to have probability zero. (The transition probabilities $t(s|s_{j-1})$ and $t(s_{j+1}|s)$ must still factor in, even if the probabilities of other transitions at position j are set to zero.) The parenthetical comments refer to a common small mistake of effectively reporting a conditional probability $p(x_1, x_2, ..., x_m, s_1, s_2, ...s_m|s_j = s)$ rather than the requested joint probability.

6 Studiousness

(a) We wish to model the grade of a student in a class that has four quizzes. For $t = 1, 2, 3, 4$, the variable S_t has value 1 if the student studied for quiz t and G_t indicates the student's cumulative grade after quiz t; it can have values A, B, C, or F. Assume that studying for quiz

t only directly affects the student's grade for that quiz. Notice that the cumulative grade through week t does depend on the grade through week $t-1$. Draw a graphical model relating $S_1, \ldots, S_4$ and $G_0, \ldots, G_4$ (with G_0 denoting the grade before taking any quizzes).

Solution:



(b) For the four expressions below, draw a line connecting any pair of expressions that are equal.

$$\Pr(G_1 \mid S_4 = 1) \qquad \Pr(G_1)$$

$$\Pr(G_1 \mid S_4 = 1, G_4 = A) \qquad \Pr(G_1 \mid G_4 = A)$$

Solution: $\Pr(G_1 \mid S_4 = 1) = \Pr(G_1)$ since G_1 and S_4 are independent. The rest are different since due to explaining away, conditioning on G_4 results in G_1 and S_4 being dependent.

(c) Suppose you want to make a maximum likelihood estimate of $P(G_t \mid G_{t-1}, S_{t-1})$ from data. Can you use EM to estimate it from a sequence $S_1, ..., S_k$, but without any samples of G_i? If so, briefly outline the method you would use. If not, explain why not.

Solution: No. The G variables have no (causal) influence over the S variables. For example, given only S data, the following CPTs are equally likely: G depends entirely on the previous G (studying is useless); and G depends entirely on the current S (e.g., subject matter has changed entirely).

(d) Let's try to estimate the parameters of this model from a sequence $G_1, ..., G_k$ using EM. For simplicity, let's consider a reduced model in which there are only two grades, A and F. Our initial model θ_0 is

G_{t-1}	S_t	$P(G_t = A)$
A	1	.6
A	0	.4
F	1	.4
F	0	.6

$P(S_t = 1)$
0.7

Assume our data (sequences of G's) is (starting with G_0): A, A, F, F, F, A, A.

What is the numerical value of $P(S_1 \mid G_0, \ldots, G_6, \theta_0)$?

Solution:

$$
\begin{aligned}
&P(S_1 = 1 \mid G_0, \ldots, G_6, \theta_0) \\
&= P(S_1 = 1 \mid G_0, G_1) \\
&= P(G_1 \mid S_1 = 1, G_0) P(S_1 = 1 \mid G_0) / P(G_1 \mid G_0) \\
&= \frac{P(G_1 = A \mid S_1 = 1, G_0 = A) P(S_1 = 1)}{P(G_1 = A \mid S_1 = 1, G_0 = A) P(S_1 = 1) + P(G_1 = A \mid S_1 = 0, G_0 = A) P(S_1 = 0)} \\
&= \frac{.6 \cdot .7}{.6 \cdot .7 + .4 \cdot .3}
\end{aligned}
$$

Values are .777777 and .22222222

(e) If the values of $P(S_i = 1 \mid D, \theta_0)$ instead were .6, .2, .4, .4, .8, .6, what would the new estimated value for $P(G_t = A | G_{t-1} = A, S_t = 1)$ in θ_1 be? (Assume we are using maximum likelihood estimation without any Laplace-type correction).

Solution: (.6 + .6) / (.6 + .2 + .6) = 1.2 / 1.4 = 6/7

(f) If the values of $P(S_i = 1 \mid D, \theta_0)$ instead were .6, .2, .4, .4, .8, .6, what would the new estimated value for $P(S_t = 1)$ in θ_1 be? (Assume we are using maximum likelihood estimation without any Laplace-type correction).

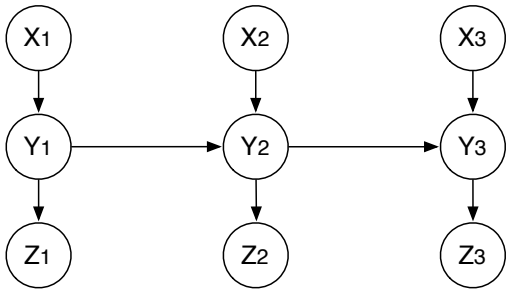
Solution: 1/2

(g) Jody suggests that, given just data G_1, ..., G_k, we might as well use a simpler model that does not have any S variables. Describe a situation, if any, in which this simpler model would be better. Describe a situation, if any, in which this simpler model would be worse.

Solution: If studying doesn't actually influence the student's grades, then a model without the S would work just as well, but otherwise we might expect it to be much worse.

7 IO HMM

Consider a graphical model, which is an HMM, but with "input" nodes $X_1, \ldots, X_t$ that are observable. So the model looks like this:



The X_t and the Z_t are observed, but the Y_t are hidden.

(a) Draw the factor graph corresponding to this model.

Solution: Make each edge undirected and insert a factor on each edge.

(b) Is this model equivalent to one in which the arrows from X_t to Y_t are reversed (allowing for different parameters)? Explain.

Solution: No. The v-structures $Y_{(t-1)} \rightarrow Y_t \leftarrow X_t$ are lost in the reversed case. In the original model, the previous hidden state $Y_{(t-1)}$ can be coupled with the input X_t to produce the current hidden state Y_t.

8 Partially Observed Markov Chain

Consider a Markov chain in which only the states on the even time steps are observed. The states are drawn from the set $\{A, B\}$.

(a) Assume we know the transition model to be:

	$\Pr(X^{t+1} = A \mid X^t)$	$\Pr(X^{t+1} = B \mid X^t)$
$X^t = A$	0.2	0.8
$X^t = B$	0.9	0.1

i. What is $\Pr(X^1 \mid X^0 = A)$?

Solution:

$$\Pr(X^1 \mid X^0 = A) = \begin{cases} 0.2 & (X^1 = A), \\ 0.8 & (X^1 = B). \end{cases}$$

ii. What is $\Pr(X^1 \mid X^0 = A, X^2 = A)$?

Solution: Using Bayes rule, we get

$$\Pr(X^1 \mid X^0 = A, X^2 = A) = \frac{1}{Z} \Pr(X^1 \mid X^0 = A) \Pr(X^2 = A | X^1).$$

Here, Z is a normalization constant, and

$$\Pr(X^1|X^0 = A)\Pr(X^2 = A|X^1) = \begin{cases} 0.2 \cdot 0.2 = 0.04 & (X^1 = A), \\ 0.8 \cdot 0.9 = 0.72 & (X^1 = B). \end{cases}$$

Therefore,

$$\Pr(X^1 \mid X^0 = A, X^2 = A) = \begin{cases} 0.04/0.76 = 0.0526 & (X^1 = A), \\ 0.72/0.76 = 0.9474 & (X^1 = B). \end{cases}$$

iii. What is $\Pr(X^1 \mid X^0 = A, X^2 = A, X^4 = A)$?

Solution: Due to the structure of the Markov chain, we have $X^1 \perp X^4 \mid X^2$. Therefore,

$$\Pr(X^1 \mid X^0 = A, X^2 = A, X^4 = A) = \Pr(X^1 \mid X^0 = A, X^2 = A).$$

This is the same as above.

(b) Now, assume we don't know the transition model, and want to estimate it from the following *observed* data sequence (note that this sequence is $\langle X^0, X^2, X^4, X^6 \rangle$)

$$D = \langle A, A, B, B \rangle$$

using the EM algorithm.

i. What are the hidden variables?

Solution: The hidden variables are X^1, X^3, and X^5.

ii. Write an expression for the log likelihood of the data, in terms of θ_{AA}, θ_{AB}, and the hidden variables. (For notational compactness, use θ_{AA} to mean $\Pr(X^{t+1} = A \mid X^t = A)$ and θ_{AB} to mean $\Pr(X^{t+1} = A \mid X^t = B)$ and note that $\theta_{BA} = 1 - \theta_{AA}$ and $\theta_{BB} = 1 - \theta_{AB}$.)

Solution: The data likelihood is:

$$\Pr(X^0 = A, X^2 = A, X^4 = B, X^6 = B; \theta) = \sum_{x^1, x^3, x^5} \Pr(X^0 = A, X^1 = x^1, X^2 = A, \ldots, X^6 = B; \theta)$$

where we are summing over all the combinations of values of the hidden variables. For our chain model, we can factor this to get

$$\sum_{x^1, x^3, x^5} \Pr(X^1 = x_1 \mid X^0 = A)\Pr(X^2 = A \mid X^1 = x^1) \Pr(X^3 = x^3 \mid X^2 = A)\Pr(X^4 = B \mid X^3 = x^3) \Pr(X^5 = x^5 \mid X^4 = B)\Pr(X^6 = B \mid X^5 = x^5).$$

Factoring, we get

$$\left(\sum_{x^1} \Pr(X^1 = x^1 \mid X^0 = A) \Pr(X^2 = A \mid X^1 = x^1)\right) \cdot$$
$$\left(\sum_{x^3} \Pr(X^3 = x^3 \mid X^2 = A) \Pr(X^4 = B \mid X^3 = x^3)\right) \cdot$$
$$\left(\sum_{x^5} \Pr(X^5 = x^5 \mid X^4 = B) \Pr(X^6 = B \mid X^5 = x^5)\right)$$

When $x^1 = A$, for example, we see that:

$$\Pr(X^1 = x^1 \mid X^0 = A) \Pr(X^2 = A \mid X^1 = x^1) = \theta_{AA}\theta_{AA}$$

Note that each combination of the hidden variables for which $X^1 = A$ contributes these same two terms to the overall likelihood. Similarly, when $x^1 = B$, we get:

$$\Pr(X^1 = x^1 \mid X^0 = A) \Pr(X^2 = A \mid X^1 = x^1) = \theta_{BA}\theta_{AB}$$

Note that we have to sum over the values of each of the variables, so, we can write an expression for the data log likelihood as:

$$\log\left((\theta_{AA}\theta_{AA} + \theta_{BA}\theta_{AB}) \cdot (\theta_{AA}\theta_{BA} + \theta_{BA}\theta_{BB}) \cdot (\theta_{AB}\theta_{BA} + \theta_{BB}\theta_{BB})\right)$$

which decomposes into

$$\log(\theta_{AA}\theta_{AA} + \theta_{BA}\theta_{AB}) + \log(\theta_{AA}\theta_{BA} + \theta_{BA}\theta_{BB}) + \log(\theta_{AB}\theta_{BA} + \theta_{BB}\theta_{BB})$$

Unfortunately, we cannot simplify this any further. **Note that this is why, in EM, we do not seek to directly optimize the log likelihood, but instead optimize the lower bound $\mathcal{L}(\hat{P}, \theta)$.**

iii. (E step) Given an initial guess of the model parameters $\theta^0_{AA}, \theta^0_{AB}$, write an expression for the posterior marginal distributions over the hidden variables conditioned on $\theta^0_{AA}, \theta^0_{AB}$ and D.

Solution: Conditioned on X^0, X^2, X^4, X^6 and model parameters, X^1, X^3, and X^5 are independent. The posterior distribution can thus be written as

$$q(X^1, X^3, X^5) = q_1(X^1)q_3(X^3)q_5(X^5).$$

For each $i = 0, 1, 2$, we have

$$q_{2i+1}(X^{2i+1} = A) = \frac{\Pr(A|X^{2i}) \Pr(X^{2i+2}|A)}{\Pr(A|X^{2i}) \Pr(X^{2i+2}|A) + \Pr(B|X^{2i}) \Pr(X^{2i+2}|B)}.$$

and

$$q_{2i+1}(X^{2i+1} = B) = 1 - q_{2i+1}(X^{2i+1} = A).$$

Note that the value for the states with even indices are given and fixed. Therefore, this can be evaluated. Particularly, we have

$$q^1(A) = \frac{\theta_{AA}\theta_{AA}}{\theta_{AA}\theta_{AA} + \theta_{BA}\theta_{AB}}, \qquad q^3(A) = \frac{\theta_{AA}\theta_{BA}}{\theta_{AA}\theta_{BA} + \theta_{BA}\theta_{BB}},$$
$$q^5(A) = \frac{\theta_{AB}\theta_{BA}}{\theta_{AB}\theta_{BA} + \theta_{BB}\theta_{BB}},$$

iv. (M step) Given D and marginal distributions on the hidden variables, what is the next estimate of the parameters, $\theta^1_{AA}, \theta^1_{AB}$? (You do not need to derive it if you can just write it down).

Solution: Remember that we approach this by optimizing $\mathcal{L}(\hat{P}, \theta)$ with respect to θ, where $\hat{P}$ is the distribution over the hidden variables, Z, given the old parameters. To

do this, we want to find the max, over θ, of

$$
\begin{aligned}
&\sum_z \hat{P}(z) \Pr(x, z; \theta) \\
&= \sum_{x_1,x_3,x_5} \hat{P}(x_1, x_3, x_5) \log \Pr(A, x_1, A, x_3, B, x_5, B; \theta) \\
&= \sum_{x_1,x_3,x_5} \hat{P}(x_1, x_3, x_5) \log(\theta_{AA}^{I(x_1=A)} \theta_{BA}^{I(x_1=B)} \theta_{AA}^{I(x_1=A)} \theta_{AB}^{I(x_1=B)} \cdot \\
&\qquad \theta_{AA}^{I(x_3=A)} \theta_{BA}^{I(x_3=B)} \theta_{BA}^{I(x_3=A)} \theta_{BB}^{I(x_3=B)} \cdot \\
&\qquad \theta_{AB}^{I(x_5=A)} \theta_{BB}^{I(x_5=B)} \theta_{BA}^{I(x_5=A)} \theta_{BB}^{I(x_5=B)}) \\
&= \sum_{x_1,x_3,x_5} \hat{P}(x_1)\hat{P}(x_3)\hat{P}(x_5)(I(x_1 = A)(l\theta_{AA} + l\theta_{AA}) + \\
&\qquad I(x_1 = B)(l\theta_{BA} + l\theta_{AB}) + \\
&\qquad I(x_3 = A)(l\theta_{AA} + l\theta_{BA}) + \\
&\qquad I(x_3 = B)(l\theta_{BA} + l\theta_{BB}) + \\
&\qquad I(x_5 = A)(l\theta_{AB} + l\theta_{BA}) + \\
&\qquad I(x_5 = B)(l\theta_{BB} + l\theta_{BB})) \\
&= \sum_{x1} \hat{P}(x_1)(I(x_1 = A)(l\theta_{AA} + l\theta_{AA}) + I(x_1 = B)(l\theta_{BA} + l\theta_{AB})) + \\
&\quad \sum_{x3} \hat{P}(x_3)(I(x_3 = A)(l\theta_{AA} + l\theta_{BA}) + I(x_3 = B)(l\theta_{BA} + l\theta_{BB})) + \\
&\quad \sum_{x5} \hat{P}(x_5)(I(x_5 = A)(l\theta_{AB} + l\theta_{BA}) + I(x_5 = B)(l\theta_{BB} + l\theta_{BB})) \\
&= \hat{P}(x_1 = A)(l\theta_{AA} + l\theta_{AA}) + \hat{P}(x_1 = B)(l\theta_{BA} + l\theta_{AB}) + \\
&\quad \hat{P}(x_3 = A)(l\theta_{AA} + l\theta_{BA}) + \hat{P}(x_3 = B)(l\theta_{BA} + l\theta_{BB})) + \\
&\quad \hat{P}(x_5 = A)(l\theta_{AB} + l\theta_{BA}) + \hat{P}(x_5 = B)(l\theta_{BB} + l\theta_{BB}) \\
&= q_1(A)(l\theta_{AA} + l\theta_{AA}) + q_1(B)(l\theta_{BA} + l\theta_{AB}) + \\
&\quad q_3(A)(l\theta_{AA} + l\theta_{BA}) + q_3(B)(l\theta_{BA} + l\theta_{BB})) + \\
&\quad q_5(A)(l\theta_{AB} + l\theta_{BA}) + q_5(B)(l\theta_{BB} + l\theta_{BB})
\end{aligned}
$$

where $l\theta_{ij}$ is shorthand for $\log \theta_{ij}$ and $I(\cdot)$ is an indicator function that is 1 when its argument is true and 0 otherwise.

In the M-step, in general, we need to optimize this objective subject to the constraints that $\theta_{AA} + \theta_{BA} = 1$, $\theta_{AB} + \theta_{BB} = 1$, and for all i, j, $\theta_{ij} \geqslant 0$. In this simple case, we can substitute $\theta_{BA} = 1 - \theta_{AA}$ and $\theta_{BB} = 1 - \theta_{AB}$ in the above expression, then set

the derivatives with respect to θ_{AA} and θ_{AB} to zero:

$$\begin{aligned}\frac{\partial}{\theta_{AA}} \quad & q_1(A)(2l\theta_{AA}) + q_1(B)(l(1-\theta_{AA}) + l\theta_{AB}) + \\ & q_3(A)(l\theta_{AA} + l(1-\theta_{AA})) + q_3(B)(l(1-\theta_{AA}) + l(1-l\theta_{AB})) + \\ & q_5(A)(l\theta_{AB} + l(1-\theta_{AA})) + q_5(B)(2l(1-\theta_{BA})) \\ = \quad & \frac{2q_1(A)}{\theta_{AA}} - \frac{q_1(B)}{1-\theta_{AA}} + \frac{q_3(A)}{\theta_{AA}} - \frac{q_3(A)}{1-\theta_{AA}} - \frac{q_3(B)}{1-\theta_{AA}} - \frac{q_5(A)}{1-\theta_{AA}} \\ = \quad & 0.\end{aligned}$$

Therefore (using $q^i(A) + q^i(B) = 1$), we get

$$\hat{\theta}_{AA} = \frac{2q^1(A) + q^3(A)}{1 + q^1(A) + 1 + q^3(A) + q^5(A)}.$$

In HMMs, the solution to this it works out that you can use "expected counts" in the regular maximum likelihood estimators. So, for example,

$$\hat{\theta}_{AA} = \frac{\text{Expected number of times A was followed by A}}{\text{Expected number of times A was followed by (A or B)}}.$$

A could have been followed by A if:

- $X_0 = A, X_1 = A$: prob is $q^1(A)$
- $X_1 = A, X_2 = A$: prob is $q^1(A)$
- $X_2 = A, X_3 = A$: prob is $q^3(A)$

A could have been followed by (A or B) in all the cases above, plus

- $X_0 = A, X_1 = B$: prob is $q^1(B) = 1 - q^1(A)$
- $X_2 = A, X_3 = B$: prob is $q^3(B) = 1 - q^3(A)$
- $X_3 = A, X_4 = B$: prob is $q^3(A)$
- $X_5 = A, X_6 = B$: prob is $q^5(A)$

Which leads to the same expression for $\hat{\theta}_{AA}$ that we found by optimization. Similar reasoning gives us $\hat{\theta}_{AB}$.

$$\hat{\theta}_{AB} = \frac{q^3(B) + q^5(A)}{q^1(B) + q^3(B) + 1 + q^5(B)}.$$