Lecture 14
Inference in Hidden Markov Models Part 1

Alice Gao

June 24, 2021

Contents

Learning Goals

By the end of the lecture, you should be able to

- Define what is the Markov assumption made by a first-order Markov chain and give some intuitions about this assumption.
- Define what is a stationary process and describe some advantages of choosing a stationary process as the model.
- Define a sensor Markov assumption.
- Describe the hidden Markov model for the umbrella world.
- Describe the important components of a hidden Markov model, why it's called hidden, and why it's called a Markov model.
- For each day in the umbrella model, derive an expression for the filtering tasks for that day.
- Explain the justification for each step in this derivation and how we obtain each probability in the final expression of the derivation.
- Given a hidden Markov model, calculate the result of filtering for Day t given the result of filtering for day $t-1$.
- Explain how to perform filtering through forward recursion.

1 Introduction

So far, we have looked at algorithms for reasoning in a static world. However, the real world changes over time. When the world changes, we need to reason about a sequence of events. In this lecture, I will introduce hidden Markov models and describe how we can use hidden Markov models to model a changing world. Hidden Markov models have many real-world applications.

2 A Hidden Markov Model for the Umbrella Story

I will use an Umbrella Story as a running example. Here's the story.

Example: You are a security guard stationed at a secret underground installation. Every day, you want to know whether it's raining or not. Unfortunately, your only access to the outside world is when you see the director brings or does not bring an umbrella each morning.

This story is a bit depressing, but it has some important elements. We are underground, but we want to know whether it's raining or not. The state of the world is whether it's raining or not, and we cannot observe it directly. Instead, we will observe a signal — whether the director comes with an umbrella or not. This signal tells us some information about the state of the world.

Also, the signal is noisy since we are in a world with uncertainty. Even if it's raining, the director might not carry an umbrella because they forget. Or if it's not raining, the director might carry an umbrella because they forgot to check the weather report.

In short, the state is not observable, but we observe a noisy signal, which tells us something about the state.

2.1 Defining Variables

Let's model the umbrella story using a Bayesian network.

We need to reason about events over time. Let's define the time steps. For the umbrella story, a day is a reasonable time step. Every day we observe a new signal, and we may want to update our estimate of the state.

Next, let's define some random variables. We need two types of random variables.

First, we need to model the state — whether it's raining or not. Let's define a binary random variable S to denote the state. S is true when it's raining and false otherwise.

We also need to model our noisy signal or observation. Let's define a binary random variable O to denote the signal. O is true when the director brings an umbrella and false otherwise.

Finally, I will use subscripts to indicate the time step for each variable.

To summarize,

- S_t is true if it rains on day t and false otherwise.
- O_t is true if the director brings an umbrella on day t and false otherwise.

2.2 The Transition Model

Next, let's construct the transition model for the umbrella story.

We need to ask the following question: How does the state change from one day to the next? How does the state today depend on the states in the past?

We're reasoning about events over time. There is one state for every time step. In general, the current state may depend on all the past states. Mathematically, we can express this as a conditional probability distribution $P(S_t|S_0 \wedge \cdots \wedge S_{t-1})$.

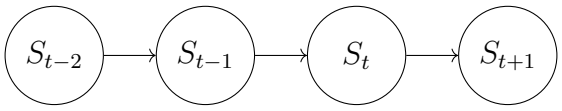
Unfortunately, this model has a significant problem. As we advance in time, the size of the conditional distribution increases. If we are modeling an arbitrary time step in the future, the conditional distribution will be un-boundedly large. An unbounded distribution is problematic since we have to store this table somewhere to perform inference.

Let's solve this problem by changing our assumption. Instead of assuming that each state depends on all of the past states, we will assume that each state depends on a fixed number of past states.

2.2.1 K-order Markov chain

Using our new assumption, we can define a K-order Markov chain. Each state depends on the K previous states.

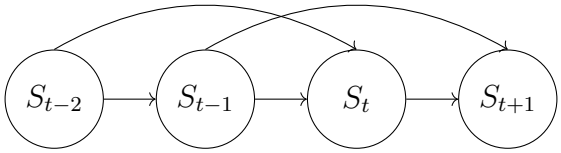
The simplest case is a first-order Markov process. Each state depends on the previous state only. Mathematically, the original transition probability is equal to the conditional probability $P(S_t|S_{t-1})$. Graphically, this model is a single chain.



The transition model:

$$P(S_t|S_{t-1} \wedge S_{t-2} \wedge S_{t-3} \wedge \cdots \wedge S_0) = P(S_t|S_{t-1})$$

In some cases, we may not be happy that each state only depends on one previous state. Perhaps, the two previous states both have useful information to determine the current state. We can define a second-order Markov process. Each state depends on the two previous states. For example, S_t depends on S_{t-1} and S_{t-2}.

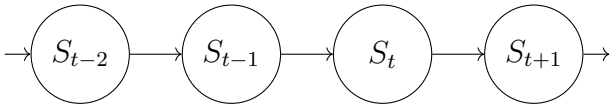


The transition model:

$$P(S_t|S_{t-1} \wedge S_{t-2} \wedge S_{t-3} \wedge \cdots \wedge S_0) = P(S_t|S_{t-1} \wedge S_{t-2})$$

We can generalize this to any fixed value of K. In a K-order Markov chain, each state depends on the previous K states.

Let's model our umbrella story as a first-order Markov process. This model makes a key assumption called the Markov assumption.



The transition model:

$$P(S_t|S_{t-1} \wedge S_{t-2} \wedge S_{t-3} \wedge \cdots \wedge S_0) = P(S_t|S_{t-1})$$

The Markov assumption says that: the current state has sufficient information to determine the next state. We do not have to look at older states in the past. I've always remembered this assumption using this sentence: "the future is independent of the past given the present."

Would you want to live in a world with the Markov assumption? I certainly would. Living in a Markovian world is wonderful because our slate gets wiped clean every day. Every day is a new beginning, and we can start fresh. Forget about the past. We need to seize the moment and do the best we can for today. What happens today determines what will happen tomorrow.

2.2.2 Stationary Process

Given the Markov assumption, how many conditional probability tables do we need to specify the transition model? In general, the transition probabilities at each time step may be different. We potentially need a separate table for each time step.

To simplify our model, we can choose to make it stationary. A stationary process doesn't mean the world does not change over time. The world still changes from one time step to the next. The word "stationary" means that how the world changes remain fixed. In other words, the transition probabilities are the same for every time step.

There are several advantages to using a stationary model. First, it's simple to specify. We can specify one conditional probability table and use it for every time step. In other words, a stationary model allows us to use a finite number of parameters to define an infinite network.

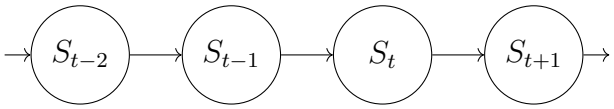
Our umbrella story can run for an unlimited number of time steps, but we can model it using a finite number of probabilities.

Second, a stationary model is a natural choice. When we're modeling real things, the dynamics often do not change. You might be wondering: what if I encounter a situation where the dynamics change? In that case, there is probably another feature causing the dynamics to change. If we model this feature explicitly, the dynamics in the new model become fixed again.

This is a partial model for the umbrella story. The states form a first-order Markov chain. The transition model has a single conditional probability table for every time step. We also have a prior distribution for the state at time 0.

$$P(s_0) = 0.5$$

$$P(s_t|s_{t-1}) = 0.7$$
$$P(s_t|\neg s_{t-1}) = 0.3$$



2.3 Sensor Model

By sensors, I mean the noisy signal we observe about the state at each time step. The signal on a day is true if the director carries an umbrella and false otherwise.

The signal at time t corresponds to the evidence variable O_t. The value of O_t could potentially depend on all the states S_0 up to S_t and all the previous signals O_0 up to O_{t-1}. As you can see, we will encounter a similar problem as before: the conditional probability distribution will grow unboundedly as time goes on.

Let's make another Markov assumption for the signals. To distinguish this from the Markov assumption regarding the states, let's call it the sensor Markov assumption. The sensor Markov assumption says that each state has sufficient information to generate its observation. Therefore, O_t needs to condition on S_t only. We can simplify the conditional probability to $P(O_t|S_t)$.

Similar to the transition model, we will assume that the sensor model is stationary. That is, the sensor model for every time step remains the same.

Below is the complete hidden Markov model for the umbrella story. Let me use this example to describe some important components of a hidden Markov model.

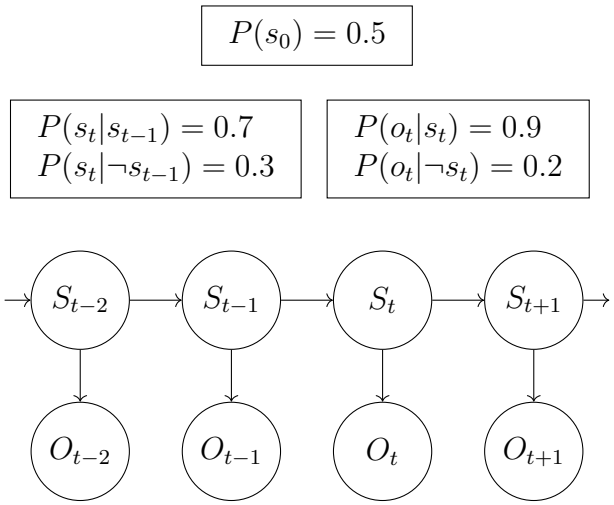
(1) The variables: The state at each time step is not observable, but the state generates a noisy signal that is observable.

(2) The state transitions satisfy the Markov assumption. Each state depends on the previous state only. Also, the sensor model satisfies the Markov assumption. Each observation depends on the current state only.

In addition, we simplified our umbrella model by making some further assumptions.

We assumed that the state transitions are stationary. The transition probabilities for each time step are the same.

We also assumed that the sensor model is stationary. The probabilities of generating a signal at each time step are the same.



3 Inference in Hidden Markov Models

3.1 Four Common Inference Tasks

There are four common inference tasks for a hidden Markov model: filtering, prediction, smoothing, and most likely explanation.

Filtering cares about what's happening today. Given the observations until today, what is the probability that I am in a particular state today? Mathematically, given the observations from day 0 to day t, we want to calculate a posterior distribution over the state on day t.

$$P(S_t|o_{0:t})$$

I am using the notation $o_{0:t}$ to represent the sequence of observations from day 0 to t. This is equivalent to $o_0 \wedge o_1 \wedge \cdots o_{t-1} \wedge o_t$. For example, we may want to estimate the probability of being in a state on day 9 given the observations from day 0 to 9.

Prediction cares about a future date. Given the observations until today, what is the probability that I am in a particular state on a day in the future? Mathematically, given the observations from day 0 to day t, we want to calculate the posterior distribution over the state on day k where k is greater than t.

$$P(S_k|o_{0:t}), k > t$$

For example, we may want to estimate the probability of being in a state on day 15 given the observations from day 0 to 9.

Smoothing cares about a day in the past. Given the observations until today, what is the probability that I was in a particular state on a day in the past? Mathematically, given the observations from day 0 to day t, we want to calculate the posterior distribution over the state on day k where $0 \leq k < t$.

$$P(S_k|o_{0:t}), 0 \leq k < t$$

For example, we may want to estimate the probability of being in a state on day 5 given the observations from day 0 to 9.

For **the most likely explanation**, we ask the following question: which sequence of states is the most likely one given our observations until today?

Among the four tasks, smoothing may be the most counter-intuitive one. You might be wondering: why do we want to perform smoothing at all? As we progress in time, we could perform filtering at every step and derive an estimate for the state at every time step. Isn't this sufficient? Think about this for a few seconds. Then, keep reading.

The main reason of performing smoothing is that making new observations can give us more information about past states. Although we performed filtering for every state in the past, since we have made new observations, our estimates are no longer accurate given the new information. We should update our estimates for all the past states given the new observations.

3.2 Algorithms for the Inference Tasks

Next, let me discuss the algorithms for performing inference. First, remember that a hidden Markov model is still a Bayesian network. Therefore, we can perform all four inference tasks using the variable elimination algorithm.

However, a hidden Markov model is a special type of Bayesian network. The particular structure of a hidden Markov model allows us to come up with specialized algorithms for performing inference. For hidden Markov models, these specialized algorithms are more efficient than the variable elimination algorithm.

There are two main algorithms. We can use the forward-backward algorithm to perform filtering and smoothing. We can use the Viterbi algorithm to derive the most likely explanation.

4 Filtering

Recall the definition of the filtering task. Let's number the time steps starting from 0. Suppose that it is day k today. Given the observations from day 0 to k, what is the probability that I am in a particular state today? Mathematically, what is

$$P(S_k|o_{0:k})?$$

$o_{0:k}$ represents a sequence of observations from day 0 to k. The small o means that we observe these signals. The capital S means that we do not observe the state.

Note that S_k can be true or false. Therefore, this quantity is not a single probability. It's a distribution containing two probabilities: the probability of $S_k = true$ given the observations and the probability of $S_k = false$ given the observations. I've written the two probabilities in angle brackets to remind you that they form a distribution.

4.1 The Forward Recursion Formulas

We can perform filtering efficiently through forward recursion. Forward recursion allows us to go through the Markov chain once and calculate all the filtering probabilities along the way. Starting from time 0, the recursion passes a distribution or a message from the time step k to the next time step k+1. Let's define a notation for this message.

$$f_{0:k} = P(S_k|o_{0:k}).$$

Again, this message is not a single probability. It is a distribution containing two probabilities.

To start, we must calculate the distribution at time 0. This is the base case. We can calculate this using the Bayes' rule. Recall that the alpha in the formula is the normalization constant. In this case, it is equal to 1 over $P(o_0)$.

$$f_{0:0} = \alpha\, P(o_0|S_0)P(S_0)$$

Once we have the message $f_{0:k-1}$, then we can calculate the message $f_{0:k}$ using recursion. Let's look at the recursive case.

$$f_{0:k} = \alpha\, P(o_k|S_k) \sum_{s_{k-1}} P(S_k|s_{k-1}) f_{0:(k-1)}$$

Given the message $f_{0:k-1}$, we want to calculate the message $f_{0:k}$. There are two other quantities. $P(o_k|S_k)$ is from the sensor model. These are the observation probabilities for time k. $P(S_k|s_{k-1})$ is from the transition model. These are the probabilities of the state transition from time k-1 to time k. We have both of these from our hidden Markov model. Also, we have a summation over the two possible states at time k-1. Finally, alpha is again the normalization constant. We do not need to know what it is. It simply means that, once we derive the two values, we need to normalize them so that they are valid probabilities.

4.2 An Example of the Base Case

Problem:

Calculate $f_{0:0} = P(S_0|O_0 = t)$ for the umbrella story.

Solution: It is tricky to use the forward recursion formulas correctly. Let me work through a few calculation examples.

Here is our umbrella model again. We will need these numbers for our examples.

First, let's calculate $f_{0:0}$. This is the base case. Assume that the observation on day 0 O_0 is true. This example only requires you to apply the Bayes' rule. However, I want to go through it to show you how the normalization constant works.

We need to calculate two probabilities. Let's calculate them separately. Write down the formula when s_0 is true. Plug in the numbers.

$$P(s_0|o_0) = \alpha\, P(o_0|s_0)P(s_0) = \alpha\, 0.9 * 0.5 = \alpha\, 0.45 \tag{1}$$

We have to stop here temporarily since normalization requires us to calculate the sum of the two values. Let's calculate the other value.

$$P(\neg s_0|o_0) = \alpha\, P(o_0|\neg s_0)P(\neg s_0) = \alpha\, 0.2 * 0.5 = \alpha\, 0.1 \tag{2}$$

Now that we have both values, let's normalize them. The first probability is equal to the first value divided by the sum of the two values. The second probability is equal to 1 minus the first probability.

$$P(s_0|o_0) = 0.45/(0.45 + 0.1) = 0.818 \tag{3}$$
$$P(\neg s_0|o_0) = 1 - 0.818 = 0.182 \tag{4}$$

Once you are familiar with these calculations, you might want to calculate the values more quickly. You can do this by using the angle bracket notation to keep track of multiple values at a time. Let me show you an example.

$$P(S_0|o_0) = \alpha\, P(o_0|S_0)P(S_0) \tag{5}$$
$$= \alpha\, \langle 0.9, 0.2\rangle * \langle 0.5, 0.5\rangle \tag{6}$$
$$= \alpha\, \langle 0.45, 0.1\rangle \tag{7}$$
$$= \langle 0.818, 0.182\rangle \tag{8}$$

Let's start by writing down the Bayes' rule formula. Remember that each term is not a single probability. It is a distribution containing two probabilities.

Let's write down the first term after alpha using the angle bracket notation. This term contains two values $P(o_0|s_0)$ and $P(o_0|\neg s_0)$. Be very careful when filling in these values. It's easy to use the incorrect values. The second term also contains two values. This is the prior distribution over S_0.

Next, we need to multiply these two terms together. This is an element-wise multiplication. The first value in the result is the product of the first values in both terms.

Finally, we need to normalize them. This step is much easier to visualize since we have both values in the same place.

4.3 An Example of the Recursive Case

Let's look at an example of the recursive case. We are given the message for time 0, $f_{0:0}$. We want to calculate the distribution for time 1, $f_{0:1}$. We assume that $O_1 = t$ — the director brings the umbrella on day 1.

Let's calculate the two probabilities using the compact notation. It is a bit challenging to plug in numbers right away, so let me re-write the formula to make it clear which numbers we need to plug in.

First, let me replace the f values with the original probability notations. This makes it clear which variables are in the terms.

$$P(S_1|o_{0:1}) = \alpha P(o_1|S_1) \sum_{s_0} P(S_1|s_0)P(s_0|o_0) \tag{9}$$

Next, let's write out the two terms in the summation explicitly. The first term is for S_0 is true and the second term is for S_0 is false.

$$= \alpha P(o_1|S_1) \left(P(S_1|s_0)P(s_0|o_0) + P(S_1|\neg s_0)P(\neg s_0|o_0)\right) \tag{10}$$

Third, let's write out the two terms for S_1 explicitly. Every time S_1 appears in a term, we should write down two values in angle brackets. The first value corresponds to the case when S_1 is true and the second value corresponds to the case when S_1 is false.

$$= \alpha\langle P(o_1|s_1), P(o_1|\neg s_1)\rangle \tag{11}$$

$$* \left(\langle P(s_1|s_0), P(\neg s_1|s_0)\rangle P(s_0|o_0) + \langle P(s_1|\neg s_0), P(\neg s_1|\neg s_0)\rangle P(\neg s_0|o_0)\right) \tag{12}$$

At this point, every variable in any term is a small letter. This means that we have explicitly written out the values of all the variables. We are ready to plug in numbers.

$$
\begin{aligned}
&\alpha\langle 0.9, 0.2\rangle \left(\langle 0.7, 0.3\rangle * 0.818 + \langle 0.3, 0.7\rangle * 0.182\right) \\
&= \alpha\langle 0.9, 0.2\rangle \left(\langle 0.5726, 0.2454\rangle + \langle 0.0546, 0.1274\rangle\right) \\
&= \alpha\langle 0.9, 0.2\rangle * \langle 0.6272, 0.3728\rangle \\
&= \alpha\langle 0.56448, 0.07456\rangle \\
&= \langle 0.883, 0.117\rangle
\end{aligned}
$$

The final answers make intuitive sense. After observing the umbrella on day 0, the probability of raining on day 0 is high, about 82%. After observing the umbrella on both days 0 and 1, the probability of raining on day 1 should be even higher and it is higher, about 88%.

4.4 Filtering Formula Derivations

In the previous section, I showed you how to calculate the filtering probabilities using the recursive formulas. While performing the calculations, have you wondered how the formulas were derived in the first place? Do you believe that these formulas are correct? Let me show you the derivations of the recursive formula in this section.

The derivation requires several steps. The formulas may look intimidating at first. However, if we break down the derivation step by step, you will realize that every step can be justified by a rule that we are already familiar with.

$$
\begin{aligned}
& P(S_k|o_{0:k}) && (13)\\
&= P(S_k|o_k \wedge o_{0:(k-1)}) && (14)\\
&= \alpha P(o_k|S_k \wedge o_{0:(k-1)})P(S_k|o_{0:(k-1)}) && (15)\\
&= \alpha P(o_k|S_k)P(S_k|o_{0:(k-1)}) && (16)\\
&= \alpha P(o_k|S_k)\sum_{s_{k-1}} P(S_k \wedge s_{k-1}|o_{0:(k-1)}) && (17)\\
&= \alpha P(o_k|S_k)\sum_{s_{k-1}} P(S_k|s_{k-1} \wedge o_{0:(k-1)})P(s_{k-1}|o_{0:(k-1)}) && (18)\\
&= \alpha P(o_k|S_k)\sum_{s_{k-1}} P(S_k|s_{k-1})P(s_{k-1}|o_{0:(k-1)}) && (19)
\end{aligned}
$$

I'll explain the derivation using six questions. For each question, I will show you a step and ask you to pick the correct justification out of five options: Bayes' rule, re-writing the expression, the chain rule or the product rule, the Markov assumption, and the sum rule.

Problem: Step 1: What is the justification for the step below?

$$
\begin{aligned}
& P(S_k|o_{0:k})\\
&= P(S_k|o_k \wedge o_{0:(k-1)})
\end{aligned}
$$

(A) Bayes' rule

(B) Re-writing the expression

(C) The chain/product rule

(D) The Markov assumption

(E) The sum rule

Solution:

The correct answer is (B). We simply re-wrote the expression. Recall that $o_{0:k}$ is a sequence of observations. This step splits up the term into two parts: one part for time k only and one part for the sequence of observations from time 0 to time k-1.

Problem: Step 2: What is the justification for the step below?

$$= P(S_k|o_k \wedge o_{0:(k-1)})$$
$$= \alpha P(o_k|S_k \wedge o_{0:(k-1)})P(S_k|o_{0:(k-1)})$$

(A) Bayes' rule

(B) Re-writing the expression

(C) The chain/product rule

(D) The Markov assumption

(E) The sum rule

Solution: The correct answer is (A) Bayes's rule. It's easier to see this when you cross out $o_{0:k-1}$ since it appears in all three terms.

$$= P(S_k|o_k \wedge \cancel{o_{0:(k-1)}})$$
$$= \alpha P(o_k|S_k \wedge \cancel{o_{0:(k-1)}})P(S_k|\cancel{o_{0:(k-1)}})$$

We effectively switched the places of S_k and O_k using Bayes' rule. The reason is that our model gives us the probability of the observation given the state, but not the probability of the state given the observation. So it is more convenient if we have a probability in the form of the probability of the observation given the state.

Problem: Step 3: What is the justification for the step below?

$$= \alpha P(o_k|S_k \wedge o_{0:(k-1)})P(S_k|o_{0:(k-1)})$$
$$= \alpha P(o_k|S_k)P(S_k|o_{0:(k-1)})$$

(A) Bayes' rule

(B) Re-writing the expression

(C) The chain/product rule

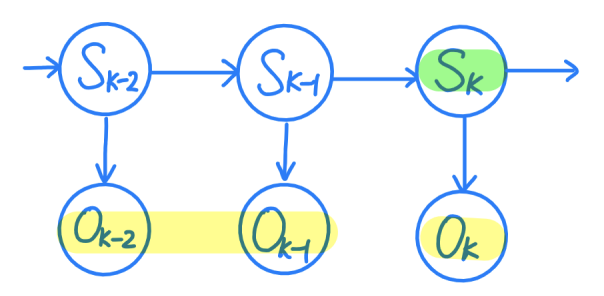
(D) The Markov assumption

(E) The sum rule

Solution: The correct answer is (D) The Markov assumption. The step removes $o_{0:k-1}$ from the first term. We can remove this value since the observation at time k only depends on the state at time k. This is the sensor Markov assumption. Given S_k, o_k is independent of any previous observations.

$$= \alpha P(o_k|S_k \wedge o_{0:(k-1)})P(S_k|o_{0:(k-1)})$$
$$= \alpha P(o_k|S_k)P(S_k|o_{0:(k-1)})$$

You can also look at the Bayesian network and verify this independence relationship using d-separation.



Problem: Step 4: What is the justification for the step below?

$$= \alpha P(o_k|S_k)P(S_k|o_{0:(k-1)})$$
$$= \alpha P(o_k|S_k)\sum_{s_{k-1}} P(S_k \wedge s_{k-1}|o_{0:(k-1)})$$

(A) Bayes' rule

(B) Re-writing the expression

(C) The chain/product rule

(D) The Markov assumption

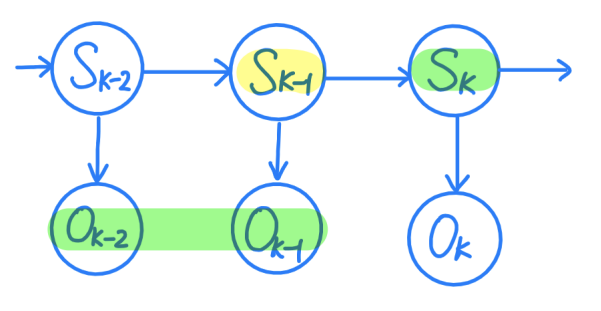
(E) The sum rule

Solution:

The correct answer is (E) The sum rule. We used the sum rule in reverse to introduce s_{k-1} into the second term.

$$= \alpha P(o_k|S_k)P(S_k|o_{0:(k-1)})$$
$$= \alpha P(o_k|S_k) \sum_{s_{k-1}} P(S_k \wedge s_{k-1}|o_{0:(k-1)})$$

We can see the reason for doing this from the Bayesian network. The second term contains S_k and $o_{0:k-1}$. These two parts are not directly connected. Introducing s_{k-1} connects S_k and $o_{0:k-1}$ in the network.



Problem: Step 5: What is the justification for the step below?

$$= \alpha P(o_k|S_k) \sum_{s_{k-1}} P(S_k \wedge s_{k-1}|o_{0:(k-1)})$$
$$= \alpha P(o_k|S_k) \sum_{s_{k-1}} P(S_k|s_{k-1} \wedge o_{0:(k-1)})P(s_{k-1}|o_{0:(k-1)})$$

(A) Bayes' rule

(B) Re-writing the expression

(C) The chain/product rule

(D) The Markov assumption

(E) The sum rule

Solution: The correct answer is (C) The chain rule or the product rule. This is easier to see if we cross out the last value $o_{0:k-1}$, which appears in all three terms. We

used the product rule to write the probability as a product of two probabilities.

$$= \alpha P(o_k|S_k) \sum_{s_{k-1}} P(S_k \wedge s_{k-1}|o_{0:(k-1)})$$
$$= \alpha P(o_k|S_k) \sum_{s_{k-1}} P(S_k|s_{k-1} \wedge o_{0:(k-1)}) P(s_{k-1}|o_{0:(k-1)})$$

Problem: Step 6: What is the justification for the step below?

$$= \alpha P(o_k|S_k) \sum_{s_{k-1}} P(S_k|s_{k-1} \wedge o_{0:(k-1)}) P(s_{k-1}|o_{0:(k-1)})$$
$$= \alpha P(o_k|S_k) \sum_{s_{k-1}} P(S_k|s_{k-1}) P(s_{k-1}|o_{0:(k-1)})$$

(A) Bayes' rule

(B) Re-writing the expression

(C) The chain/product rule

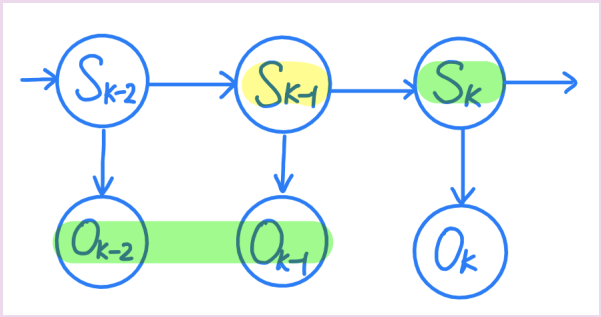
(D) The Markov assumption

(E) The sum rule

Solution: The correct answer is (D) The Markov assumption. This step removes the value $o_{0:k-1}$ from the first term. We can remove this value since the state at time k only depends on the state at time k-1. This is the Markov assumption. Given S_{k-1}, S_k is independent of any previous observations.

$$= \alpha P(o_k|S_k) \sum_{s_{k-1}} P(S_k|s_{k-1} \wedge o_{0:(k-1)}) P(s_{k-1}|o_{0:(k-1)})$$
$$= \alpha P(o_k|S_k) \sum_{s_{k-1}} P(S_k|s_{k-1}) P(s_{k-1}|o_{0:(k-1)})$$

Again, you can verify this using d-separation.


S_{k-2}
S_{k-1}
S_k
O_{k-2}
O_{k-1}
O_k