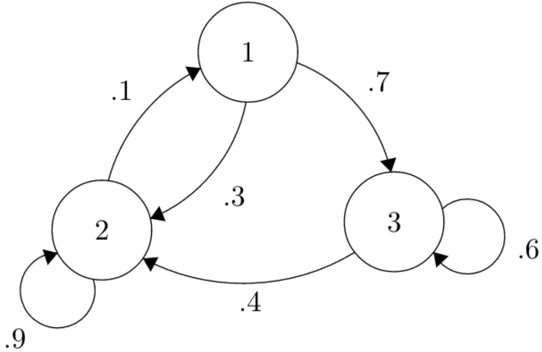
Oct 2~~6~~8 Hidden Markov Models part 2

Opening Example: *Set up* the detailed balance condition to find the stationary distribution of the network:



Announcements and To-Dos

Announcements:

1. Skip 1a for now but it's worth a bit of E.C. if you get A^* working. I'll add a few edges to hard code in an addendum.

Last time we learned:

1. Stationary distributions to Markov Models.

Stationarity Recap

Definition: We say that a Markov chain has reached its *stationary distribution* if $P(X_{t+1}) = P(X_t)$. (long run behavior)

Definition When the flow into x' from x = Flow into x from x' for *each and every* pair x, x', we say that $q(x', x)$ is in *detailed balance* with probability $\pi(x)$. This *implies* stationarity.

Definition: The transition probability distribution q is called *ergodic* if every state is *reachable* from every other state, and there are no strictly periodic cycles.

Proposition: If a Markov chain is *ergodic*, then there exists a **unique** stationary distribution for any given set of transition probabilities.

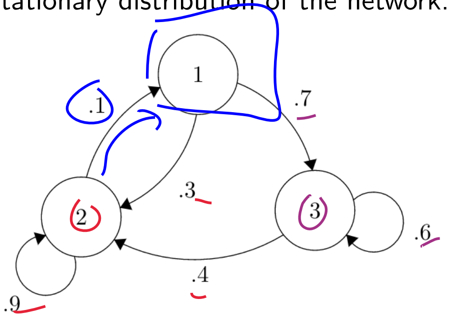
Opening Soln

$1: a + b = 1$

$2: 2a + 2b = 2$

☹

Opening Example: *Set up* the detailed balance condition to find the stationary distribution of the network:



1): In = Out

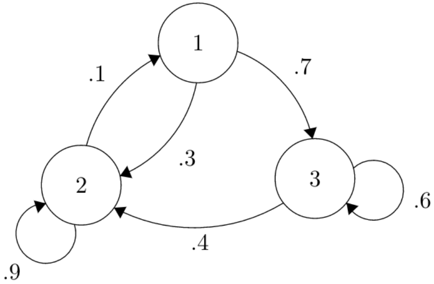
$.1 \cdot X_2 = X_1$

2) $.3 \cdot X_1 + .9 \cdot X_2 + .4 \cdot X_3 = X_2$

3) $.7 X_1 + .6 X_3 = X_3$

Opening Soln

Opening Example: *Set up* the detailed balance condition to find the stationary distribution of the network:



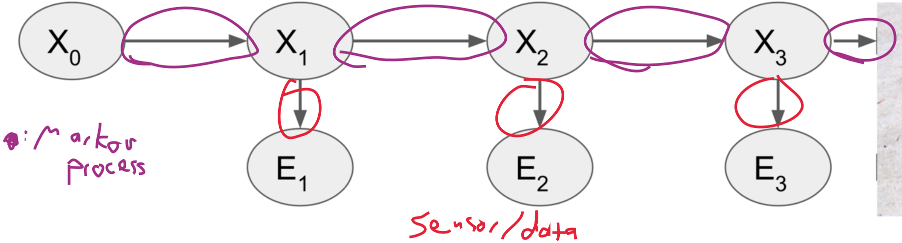
1. Node 1 in = Node 1 out: $.1X_2 = X_1$
2. Node 2 in = Node 2 out: $.3X_1 + .9X_2 + .4X_3 = X_2$
3. Node 3 in = Node 3 out: $\boxed{.7X_1 + .6X_3 = X_3}$
   Three equations, three unknowns?!...but we already *knew* that X_3 was receiving the unaccounted for $.7$ from X_1 and the last $.6$ from itself. This line is not *linearly independent* of the others, so it's actually useless.
4. Final line: $\boxed{X_1 + X_2 + X_3 = 1.}$

Hidden Markov Models

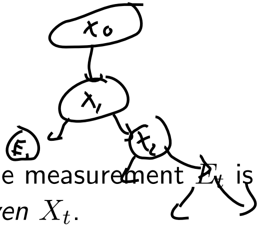
Example:

Suppose you are a graduate student in a basement office. You are writing your dissertation, so you don't get to leave very often.

You are curious if it is raining, and the only contact you have with the outside world is through your advisor. If it is raining, she brings her umbrella 90% of the time, and has it just in case on 20% of sunny days. You know that historically, 40% of rainy days were followed by another rainy day, and 30% of sunny days were followed by a rainy day.



Hidden Markov Models: Sensoring

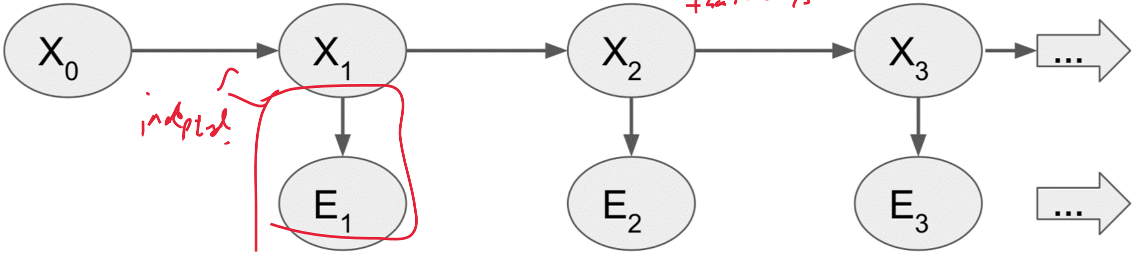


Denote $X_{0:t} = [X_0, X_1, X_2, \overset{X_t}{\dots}]$ as the states at each time step.

Definition: The *Sensor Markov Assumption* is the assumption that the measurement E_t is conditionally independent of all previous measurements and states, *given* X_t.

$$P(E_t | X_{0:t}, E_{0:t-1}) = P(E_t | X_t)$$

about evidence . . . . that day! weather



Hidden Markov Models: Roadmap

An assumption like $P(E_t|X_{0:t}, E_{0:t-1}) = P(E_t|X_t)$ is powerful, because we'll get to use large probability products and conditional probability tables just like we did with Bayesian Networks. In general, we have a handful of tasks to do on networks like this:

1. **Filtering:** Describing the *process*.
2. **Prediction:** Describing the future: X_{t+1} *given* the past.
3. **Smoothing:** Describing the past: (or the chain X). → probabilities
4. **Most likely explanation:** Describing the past: (or the chain X). → one outcome
5. **Learning:** Bayesian updates and improvements on *priors* and *posteriors*.

Hidden Markov Models: Roadmap

Any of these processes are often decomposed into the two primary tasks of statistics and data science:

Estimation:

1. Come up with a *model* for prediction and explanation - the network
2. Compare the model to data
3. e.g. $P(Alarm|MaryCalls) = ?$.

Inference:

1. *Validate* your model. What does the data tell us about the model?
2. E.g. hill-climbing or annealing for SLR parameters (...and getting an R^2!)

We will often have do both!

HMM

Any of these processes are often decomposed into the two primary tasks of statistics and data science:

Estimation:

1. Come up with a *model* for prediction and explanation
2. Compare the model to data
3. e.g. $P(Alarm|MaryCalls) = ?$.

Inference:

1. *Validate* your model. What does the data tell us about the model?
2. E.g. hill-climbing or annealing for SLR parameters (...and getting an R^2!)

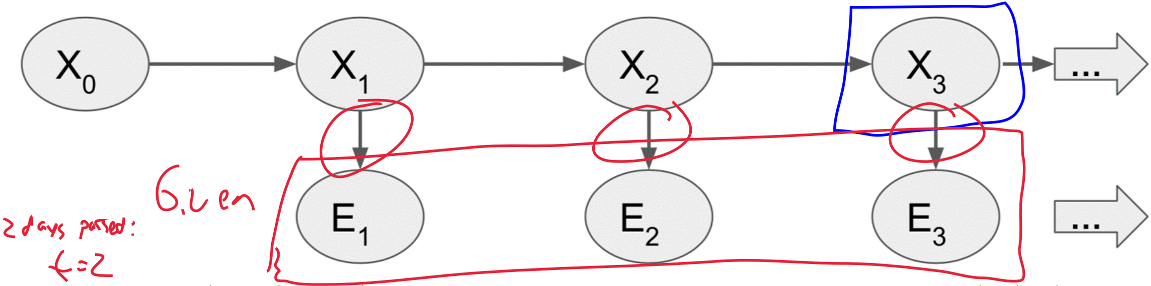
We will often have do both!

HMM: Filtering

Conditioning: $P(A,B,C) = P(A \text{ and } B \text{ and } C) = P(A \& B \mid C) \cdot P(C)$
$= P(A \mid (B \& C)) \cdot P(B \& C)$
$= P(C \mid A \& B)\, P(A \& B)$

predict

Filtering: The goal is to predict X_{t+1} *given* all the evidence available $E_{1:t+1}$.



So we want: $P(X_{t+1}|E_{1:t+1})$. It turns out, since we know the sensor model $P(E_t|X_t)$, it's worth splitting up evidence into the past and the present:

$P(ABC) = P(A \mid BC) \cdot P(B \cdot C)$

$P(X_3 \mid E_{1:3})$

$$P(X_{t+1}|E_{1:t+1}) = P(X_{t+1}|E_{1:t}, E_{t+1})$$

past info

current information

HMM: Filtering

$$P(A|BC) = \frac{P(C|AB)\cdot P(AB)}{P(BC)}$$

$$\begin{aligned} P(X_{t+1}|E_{1:t+1}) &= P(\overbrace{X_{t+1}}^{A}|\overbrace{E_{1:t}}^{B},\overbrace{E_{t+1}}^{C}) \\ &\overset{Bayes}{=} \frac{P(E_{t+1}|X_{t+1},E_{1:t})P(X_{t+1},E_{1:t})}{P(E_{1:t,t+1})} \\ &= \alpha P(E_{t+1}|X_{t+1},E_{1:t})P(X_{t+1},E_{1:t}) \end{aligned}$$

reality at time t+1 given all past evidence

evidence at time t+1 given reality at time t+1

P of today's weather and past evidence.

denom = don't care

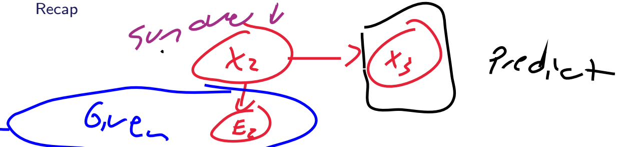
?

(pt/note)

Now, it's *also* worth splitting up X into the past and the present, since we're given the transition model for $X_{t+1}|X_t$.

So we're going to break up $X_{t+1}, E_{1:t}$ by using the Law of Total Probability and summing over all possible states of X_t, and get that

HMM: Filtering



$$\begin{aligned}
P(X_{t+1}|E_{1:t+1}) =& P(X_{t+1}|E_{1:t}, E_{t+1}) \\
\overset{Bayes}{=}& \frac{P(E_{t+1}|X_{t+1}, E_{1:t})P(X_{t+1}, E_{1:t})}{P(E_{1:t,t+1})} \\
=& \alpha P(E_{t+1}|X_{t+1}, E_{1:t})P(X_{t+1}, E_{1:t})
\end{aligned}$$

Now, it's *also* worth splitting up X into the past and the present, since we're given the transition model for $X_{t+1}|X_t$.

So we're going to break up $X_{t+1}, E_{1:t}$ by using the Law of Total Probability and summing over all possible states of X_t, and get that

Law total prob

$$\begin{aligned}
P(X_{t+1}, E_{1:t}) &= \sum_{X_t} P(X_{t+1}, E_{1:t}|X_{1:t}) \\
&= \sum_{X_t} P(X_{t+1}|E_{1:t}X_{1:t})P(X_t|E_{1:t})
\end{aligned}$$

prior X →

HMM: Filtering

We want:

denominator

evidence given reality

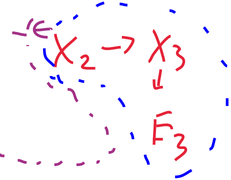
reality given past evidence

$$\boxed{P(X_{t+1}|E_{1:t+1})} = \overset{\text{denominator}}{\alpha} \overset{\text{evidence given reality}}{P(E_{t+1}|X_{t+1}, E_{1:t})} \underbrace{P(X_{t+1}, E_{1:t})}$$

$$= \alpha P(E_{t+1}|X_{t+1}, \cancel{E_{1:t}}) \underbrace{\sum_{X_t} P(X_{t+1}|E_{1:t}X_{1:t}) P(X_t|E_{1:t})}$$

Sensor and Transition Models give independence!

$$= \alpha P(E_{t+1}|X_{t+1}) \sum_{X_t} \underbrace{P(X_{t+1}|X_t)}_{\text{markov model}} P(X_t|E_{1:t})$$



That last term is the same $X|E$ as the left-hand side, but for one *prior* time step. Sounds like a recursion or induction problem!

HMM: Filtering

We want:

$$\begin{aligned}
P(X_{t+1}|E_{1:t+1}) &= \alpha P(E_{t+1}|X_{t+1}, E_{1:t})P(X_{t+1}, E_{1:t}) \\
&= P(E_{t+1}|X_{t+1}, E_{1:t}) \sum_{X_t} P(X_{t+1}|E_{1:t}X_{1:t})P(X_t|E_{1:t})
\end{aligned}$$

Sensor and Transition Models give independence!

$$\begin{aligned}
&= P(E_{t+1}|X_{t+1}) \sum_{X_t} P(X_{t+1}|X_t)P(X_t|E_{1:t}) \\
&= \underbrace{P(E_{t+1}|X_{t+1})}_{Sensor} \sum_{X_t} \underbrace{P(X_{t+1}|X_t)}_{Transition} \underbrace{P(X_t|E_{1:t})}_{One\,prior\,time\,step}
\end{aligned}$$

$\rightarrow$ start from
$t=0$, gives
$t=1$, gives
$t=2$ . . . .

That last term is the same $X|E$ as the left-hand side, but for one *prior* time step. Sounds like a recursion or induction problem!

HMM: Forward Algorithm

$$P(X_{t+1}|E_{1:t+1}) = \alpha P(E_{t+1}|X_{t+1}) \sum_{X_t} P(X_{t+1}|X_t) P(X_t|E_{1:t})$$

Where we update in the form of

$$f_{1:t+1} = \underline{\alpha \, \text{Forward}(f_{1:t}, E_{t+1}).}$$

evidence up to t+1

Consider the Umbrella/advisor example.

Sensor Model → yes umbrella

X_t	$P(E_t\|X_t)$
rain T	.9
sun F	.2

Transition Model

X_t	$P(X_{t+1}\|X_t)$
rain T	.4
sun F	.3

Suppose that we take 3 time steps, with evidence of $E_{1:3} = [T, F, T]$, and we assign a prior for X_0 of $[.5, .5]$.

HMM: Forward Algorithm

LTP: $P(A) = P(A \text{ and } B=T) + P(A \text{ and } B=F)$

$= P(A \mid B=T) \cdot P(B=T) + P(A \mid B=F) \cdot P(B=F)$

$$P(X_{t+1}|E_{1:t+1}) = \alpha P(E_{t+1}|X_{t+1}) \sum_{X_t} P(X_{t+1}|X_t) P(X_t|E_{1:t})$$

Sensor Model

X_t	$P(E_t\|X_t)$
T	.9
F	.2

Transition Model

X_t	$P(X_{t+1}\|X_t)$
T	.4
F	.3

Initializations

$E_{1:3} = [T, F, T]$,

$X_0 = [.5, .5]$

Prior

Then we update, starting at $t = 0$:

$$P(X_1 | E_{1:1}) = \alpha \, \overbrace{P(E_1 | X_1)} \sum_{\substack{X_0 = sun \\ X_0 = rain}} \boxed{P(X_1 | X_0)} \cdot P(X_0 | \cancel{E})$$

$X_0 = T$ sun

$X_0 = F$ rain

2 terms in sum

HMM: Forward Algorithm

$$P(X_{t+1}|E_{1:t+1}) = \alpha \boxed{P(E_{t+1}|X_{t+1})} \sum_{X_t} P(X_{t+1}|X_t)P(X_t|E_{1:t})$$

Sensor Model

X_t	$P(E_t\|X_t)$
T	.9
F	.2

Transition Model

X_t	$P(X_{t+1}\|X_t)$
T	.4
F	.3

$x_1 | x_0$

Initializations

$E_{1:3} = [T, F, T]$,

$X_0 = [.5, .5]$

$P(X_1=T | x_0=T) = .4$ $P(X_1=F | X_0=T) = .6$

Then we update, starting at $t = 0$:

$$P(X_1|E_1 = T) = \alpha P(E_1|X_1) \sum_{X_0} P(X_1|X_0)P(X_0|E_{1:0})$$

$\alpha = .315 + .13$

evidence =T

$$\begin{pmatrix} P(X_1 = T|E_1 = T) \\ P(X_1 = F|E_1 = T) \end{pmatrix} = \alpha \begin{pmatrix} 0.9 \\ 0.2 \end{pmatrix} \left[\underbrace{\begin{pmatrix} 0.4 \\ 0.6 \end{pmatrix} (0.5)}_{X_0=T} + \underbrace{\begin{pmatrix} 0.3 \\ 0.7 \end{pmatrix} (0.5)}_{X_0=F} \right] = \alpha \begin{pmatrix} 0.315 \\ 0.13 \end{pmatrix}$$

HMM: Forward Algorithm

$$P(X_{t+1}|E_{1:t+1}) = \alpha P(E_{t+1}|X_{t+1}) \sum_{X_t} P(X_{t+1}|X_t) P(X_t|E_{1:t})$$

$P(X_2 | E_{1:2})$

Sensor Model

X_t	$P(E_t\|X_t)$
T	.9
F	.2

Transition Model

X_t	$P(X_{t+1}\|X_t)$
T	.4
F	.3

Initializations

$E_{1:3} = [T, F, T]$,

$X_0 = [.5, .5]$,

$X_1 = [.708, .292]$

$P(X_1 | E_1)$

plug into $t=1$

Then we update again, starting at $t = 1$:

HMM: Forward Algorithm

gives us any
$P(X_{t+1} \mid \overline{E_{1:t+1}})$

$$P(X_{t+1}|E_{1:t+1}) = \alpha P(E_{t+1}|X_{t+1}) \sum_{X_t} P(X_{t+1}|X_t)P(X_t|E_{1:t})$$

Sensor Model

X_t	$P(E_t \mid X_t)$	$P(E_t=F \mid X_t)$
T	.9	.1
F	.2	.8

Transition Model

X_t	$P(X_{t+1} \mid X_t)$
T	.4
F	.3

Initializations

$E_{1:3} = [T, F, T]$,
$X_0 = [.5, .5]$,
$X_1 = [.708, .292]$

last time
$P(X_0)$

$P(X_1 \mid E_1)$

Then we update again, starting at $t = 1$:

$$P(X_2|E_1 = T,\, E_2 = F) = \alpha P(E_2|X_2) \sum_{X_1} P(X_2|X_1)P(X_1|E_{2:0})$$

evidence:
$E_2 = F$

$$= \alpha \boxed{\begin{pmatrix}0.1\\0.8\end{pmatrix}} \left[\underbrace{\begin{pmatrix}0.4\\0.6\end{pmatrix}\boxed{(0.708)}}_{X_1=T} + \underbrace{\begin{pmatrix}0.3\\0.7\end{pmatrix}\boxed{(0.292)}}_{X_1=F} \right] = \alpha_1 \begin{pmatrix}0.037\\0.503\end{pmatrix}$$

HMM: Prediction

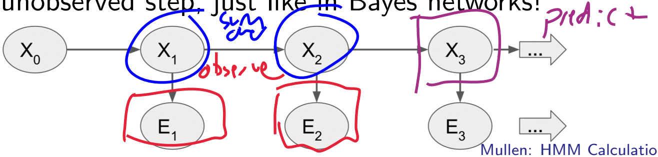
It turns out, this setup also allows us to *skip* steps and predict things in the future.

$$P(X_{t+1}|E_{1:t}) = \sum_{X_t} P(X_{t+1}|X_t)P(X_t|E_{1:t})$$

is a prediction *one* time step in the future. The same setup works if we skip ahead by k.

$$P(X_{t+k+1}|E_{1:t+1}) = \sum_{X_{t+k}} P(X_{t+k+1}|X_{t+k})P(X_{t+k}|E_{1:t})$$

This can also be done recursively, but now we have to sum over all possible outcomes of each unobserved step, just like in Bayes networks!



HMM: Prediction

So we want to predict the time in the future k...

$$P(X_{t+k+1}|E_{1:t+1}) = \sum_{X_{t+k}} P(X_{t+k+1}|X_{t+k})P(X_{t+k}|E_{1:t})$$

A one-step prediction is $k = 0$:

$$P(X_{t+1}|E_{1:t}) = \sum_{X_t} P(X_{t+1}|X_t)P(X_t|E_{1:t})$$

A one-step prediction is $k = 1$:

$$P(X_{t+2}|E_{1:t}) = \sum_{X_{t+1}} P(X_{t+1}|X_{t+1})P(X_{t+1}|E_{1:t})$$

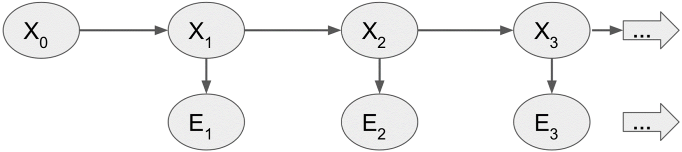
but the last term *is* the one-step prediction!

$$P(X_{t+2}|E_{1:t}) = \sum_{X_{t+1}} P(X_{t+1}|X_{t+1}) \sum_{X_t} P(X_{t+1}|X_t)P(X_t|E_{1:t})$$

HMM: Smoothing

Our final task is **smoothing**, where we try to update probabilities of prior states X based on current evidence.

So we want a description for $P(X_k|E1:t)$, where $t > k$.



$$\begin{aligned} P(X_k|E_{1:t}) &= P(X_k|E_{1:k}, E_{k+1:t}) \\ &= \alpha P(E_{k+1:t}|X_k, E_{1:k})P(X_k|E_{1:k}) \\ &= \alpha P(E_{k+1:t}|X_k)P(X_k|E_{1:k}) \end{aligned}$$

We can find the last term by the FORWARD algorithm for filtering.

HMM: Smoothing

This leaves the $P(E_{k+1:t}|X_k)$ term, which we denote by $b_{k+1:t}$, which is the probability of future *measurements* given the current state of our system, which is just the combination of our transition and sensor models! Imagine taking *one* time step and asking about the new evidence: we need to describe X_{k+1}.

$$\begin{aligned}
b_{k+1:t} &= P(E_{k+1:t}|X_k) \\
&= \sum_{X_{k+1}} P(E_{k+1:t}|\underbrace{X_k, X_{k+1}}_{indep})P(X_{k+1}, X_k) \\
&= \sum_{X_{k+1}} P(\underbrace{E_{k+1:t}}_{split\, up}|X_{k+1})P(X_{k+1}, X_k) \\
&= \sum_{X_{k+1}} P(\underbrace{E_{k+1}, E_{k+2:t}}_{conditional}|X_{k+1})P(X_{k+1}, X_k) \\
&= \sum_{X_{k+1}} \underbrace{P(E_{k+1}|X_{k+1})}_{sensor\, model}\underbrace{P(E_{k+2:t}|X_{k+1})}_{similar\, to\, LHS}\underbrace{P(X_{k+1}, X_k)}_{Markov\, model}
\end{aligned}$$

HMM: Smoothing

The middle term is a *backwards* model, since we're working from the past (oldest unobserved state) rather what FORWARD did.

$$\begin{aligned} b_{k+1:t} &= P(E_{k+1:t}|X_k) \\ &= \sum_{X_{k+1}} \underbrace{P(E_{k+1}|X_{k+1})}_{sensor\, model} P(E_{k+2:t}|X_{k+1}) \underbrace{P(X_{k+1}, X_k)}_{Markov\, model} \\ &= \textsc{Backward}(b_{k+2:t}, E_{k+1}) \end{aligned}$$

All told, then, we have:

$$\begin{aligned} P(X_k|E_{1:t}) &= P(X_k|E_{1:k}, E_{k+1:t}) \\ &= \alpha P(E_{k+1:t}|X_k, E_{1:k}) P(X_k|E_{1:k}) \\ &= \alpha P(E_{k+1:t}|X_k) P(X_k|E_{1:k}) \\ &= \alpha \textsc{Backward} \times \textsc{Forward} \\ &= \alpha(b_{k+1:t}) \times (f_{1:k}) \end{aligned}$$

HMM: Smoothing

What does this actually look like?

$$\begin{aligned} P(X_1|E_{1:3}) &= \text{What's the probability it rained on Day 1 given 3 days of evidence (TFT)?} \\ &= \alpha f_{1:1} b_{2:3} \end{aligned}$$

$$\begin{aligned} f_{1:1} &= \text{What's the probability it rained on Day 1 given evidence through day 1?} \\ &= P(X_1|E_{1:1}) = \alpha P(E_1|X_1) \sum_{X_0} P(X_1|X_0) P(X_0|E_{null}) \\ &= \begin{pmatrix} .708 \\ .292 \end{pmatrix} \end{aligned}$$

HMM: Smoothing

We have to run $\textsc{Backwards}$ for $k = 1$ and $k = 2$. ($t = 3$ for both!)

$$\begin{aligned}
b_{2:3} &= \text{What's the probability of the evidence on days 2 and 3, given X at day 2?} \\
&= P(E_{2:3}|X_2) \\
&= \sum_{X_2} P(E_2|X_2)P(E_{3:3}|X_2)P(X_2|X_1) \\
&= \sum_{X_2} P(E_2|X_2)b_{3:3}P(X_2|X_1) \\
b_{3:3} &= \text{What's the probability of the evidence on days 3-3, given X at day 2?} \\
&= P(E_{3:3}|X_2) \\
&= \sum_{X_3} P(E_3|X_3)P(E_{4:3}|X_3)P(X_3|X_2) \\
&= \sum_{X_3} P(E_3|X_3)b_{4:3}P(X_3|X_2) \text{ but } b_{4:3} = 1 \text{ by independence!}
\end{aligned}$$

HMM: Smoothing

We have to run $\textsc{Backwards}$ for $k = 1$ and $k = 2$. ($t = 3$ for both!)

$$b_{2:3} = \sum_{X_2} P(E_2|X_2) b_{3:3} P(X_2|X_1)$$

$$b_{3:3} = \sum_{X_3} P(E_3|X_3) P(X_3|X_2)$$

$$= \alpha \left[\underbrace{\begin{pmatrix} 0.4 \\ 0.3 \end{pmatrix} (0.9)}_{X_3=T} + \underbrace{\begin{pmatrix} 0.6 \\ 0.7 \end{pmatrix} (0.2)}_{X_3=F} \right] = \begin{pmatrix} 0.48 \\ 0.41 \end{pmatrix}$$

Sensor Model

X_t	$P(E_t\|X_t)$
T	.9
F	.2

Transition Model

X_t	$P(X_{t+1}\|X_t)$
T	.4
F	.3

Initializations and Evidence

$E_{1:3} = [T, F, T]$,
$X_0 = [.5, .5]$

HMM: Smoothing

We have to run $\textsc{Backwards}$ for $k = 1$ and $k = 2$. $(t = 3$ for both!)

$$b_{2:3} = \sum_{X_2} P(E_2|X_2) b_{3:3} P(X_2|X_1)$$

$$= \alpha \left[\underbrace{\begin{pmatrix} 0.4 \\ 0.3 \end{pmatrix} (0.48)\,(0.1)}_{X_2=T} + \underbrace{\begin{pmatrix} 0.6 \\ 0.7 \end{pmatrix} (0.41)\,(0.8)}_{X_2=F} \right] = \begin{pmatrix} 0.68 \\ 0.32 \end{pmatrix}$$

Sensor Model

X_t	$P(E_t\|X_t)$
T	.9
F	.2

Transition Model

X_t	$P(X_{t+1}\|X_t)$
T	.4
F	.3

Initializations and Evidence

$E_{1:3} = [T, F, T]$,
$X_0 = [.5, .5]$

HMM: Smoothing

Sensor Model

X_t	$P(E_t \mid X_t)$
T	.9
F	.2

Transition Model

X_t	$P(X_{t+1} \mid X_t)$
T	.4
F	.3

Initializations and Evidence

$E_{1:3} = [T, F, T]$,
$X_0 = [.5, .5]$

We had two calculations:

$$P(X_1|E_1) = \begin{pmatrix} 0.708 \\ 0.292 \end{pmatrix} \qquad P(X_1|E_{1:3}) = \begin{pmatrix} 0.682 \\ 0.318 \end{pmatrix}$$

Sanity check? Why is the $P(X_1 = T)$ smaller with more evidence?

HMM: Smoothing

Sensor Model

X_t	$P(E_t \mid X_t)$
T	.9
F	.2

Transition Model

X_t	$P(X_{t+1} \mid X_t)$
T	.4
F	.3

Initializations and Evidence

$E_{1:3} = [T, F, T]$,

$X_0 = [.5, .5]$

We had two calculations:

$$P(X_1|E_1) = \begin{pmatrix} 0.708 \\ 0.292 \end{pmatrix} \qquad P(X_1|E_{1:3}) = \begin{pmatrix} 0.682 \\ 0.318 \end{pmatrix}$$

Sanity check? Why is the $P(X_1 = T)$ smaller with more evidence?

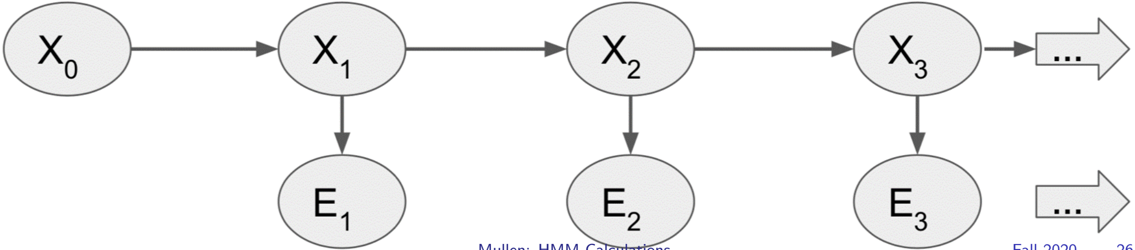
Solution: we saw evidence on rain on day 2, which in turn means it was likely to have rained on day 2... and a rainy day more likely preceded by another rainy one!

HMM: All at Once

So for any given observation X_k, we tend to have to run both a forward algorithm to ask what the evidence *up to time* k did, then a backwards algorithm to ask what the evidence afterwards did. To solve the whole chain, we do *both*. Given evidence up to time t, we:

- Run the Forward algorithm to filter it.
- then run the backward algorithm to smooth it

We use $f_{1:k}$ in the backwards algorithm, so we'll save them: the main tenet of dynamic programming is to not solve the same problem twice!



HMM: Wrapup

There's a final question that often is asked: what's the *most likely* sequence of X values that gave rise to our evidence.

- Lazy way: compute the $P(X|E)$ values and pick the most likely one for each time individually.
- Rigorous way: compute a maximization over all the nodes of $P(X_0, X_1, \ldots X_t, E_0, E_1, \ldots E_t$.

It turns out the rigorous problem can heavily exploit our independence assumptions, as usual! The joint density of the HMM will factor into

$$\Pi_{all\,nodes} P(Z_i|\text{parents}(Z_i))$$
$$= P(X_0)P(X_1|X_0)P(E_1|X_1)P(X_2|X_1)P(E_2|X_2)\ldots$$

HMM: Most Likely Sequence

$$P(X,E) = P(X_0)P(X_1|X_0)P(E_1|X_1)P(X_2|X_1)P(E_2|X_2)\dots$$

And we want to maximize this thing... maximizing products is harder than sums, so we hit with a $\log$, which keeps the max in the same place and changes products to sums.

$$\log P(X,E) = \log\left(P(X_0)P(X_1|X_0)P(E_1|X_1)\right) + \sum \log\left(P(X_k|X_{k-1})P(E_k|X_k)\right)$$

The recursive algorithm for this is called the *Viterbi* algorithm and is computed in linear time. Idea: find the best sequence for X_0, then the best sequence through time 1 including X_1, then through X_2, etc.

Moving Forward

- Coming up:
  1. Markov Decision Processes!
  2. Markov NB on Friday.