CSC384: Intro to Artificial Intelligence

Probabilistic Reasoning with Temporal Models

- This material is covered in Chapter 15 (we cover a subset of this chapter)
- Thanks to Faheim Bacchus and Peter Abbeel for slides

Uncertainty

- In many practical problems we want to reason about a sequence of observations
  - Speech recognition
  - Robot localization
  - User attention
  - Medical monitoring
- Need to introduce time (or space) into our models

Markov Models

- Say we have one variable X (perhaps with a very large number of possible value assignments).
- We want to track the probability of different values of X (i.e. the probability distribution over X) as its values change over time.
- Possible solution: Make multiple copies of X, one for each time point (we assume a discrete model of time): $X_1, X_2, X_3 \dots X_t$
- A Markov Model is specified by the two following assumptions:
  - The current state X_t is conditionally independent of the earlier states given the previous state.

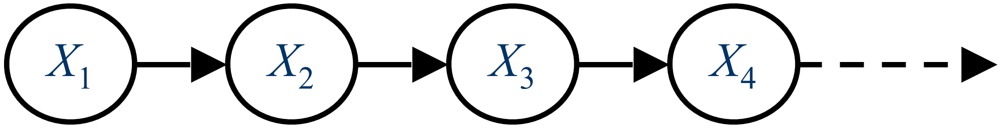
$$P(X_t \mid X_{t-1}, X_{t-2}, \dots. X_1) = P(X_t \mid X_{t-1})$$

- The transitions between X_{t-1} and X_t are determined by probabilities that do not change over time (they are stationary probabilities).

$$P(X_t \mid X_{t-1})$$

Markov Models

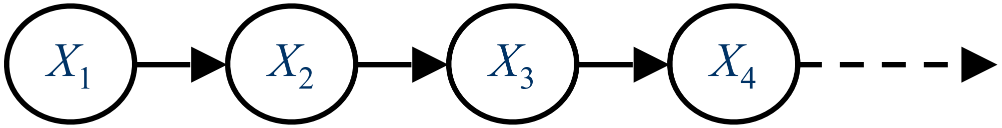
- These assumptions give rise to a Bayesian Network that looks like this:



- $P(X_1,X_2,X_3,\ldots.) = P(X_1)P(X_2|X_1)P(X_3|X_2)$ … (Assumption 1)

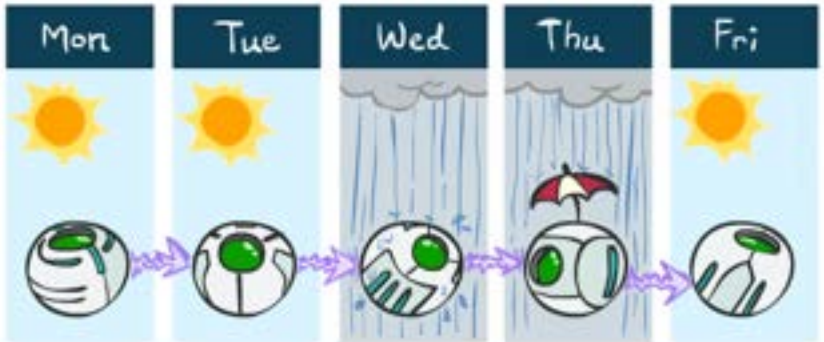
- All the CPTs (except $P(X_1)$) are the same (Assumption 2)

Markov Models



- D-Separation tells us that X_{t-1} is conditionally independent of X_{t+1}, X_{t+2}, … given X_t
  - The current state separates the past from the future.

Example Markov Chain Weather



- States: X = {rain, sun}

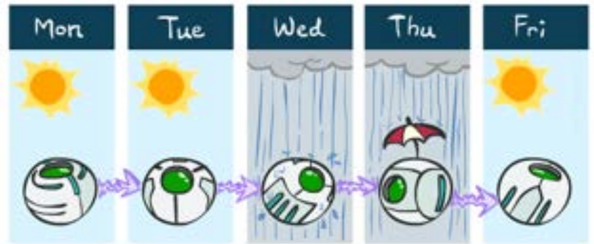
- Initial distribution:
  $P(X_1=sun) = 1.0$

CPT $P(X_t \mid X_{t-1})$:

X_{t-1}	X_t	$P(X_t \mid X_{t-1})$
sun	*sun*	*0.9*
sun	rain	0.1
rain	*sun*	*0.3*
rain	rain	0.7

[Slide created by Dan Klein and Pieter Abbeel for CS188 Intro to AI at UC Berkeley].

Example Markov Chain Weather



- $P(X_1 = \text{sun}) = 1.0$
- What is the probability distribution after one step, $P(X_2)$?
- Use summing out rule with X_1

$$P(X_2 = \text{sun}) = P(X_2 = \text{sun} | X_1 = \text{sun}) P(X_1 = \text{sun}) + P(X_2 = \text{sun} | X_1 = \text{rain}) P(X_1 = \text{rain})$$

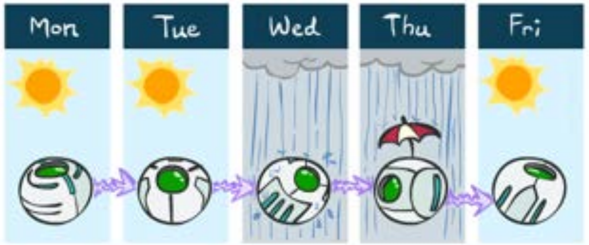
$$0.9 \cdot 1.0 + 0.3 \cdot 0.0 = 0.9$$

CPT P(Xt | Xt-1):

X_{t-1}	X_t	$P(X_t \| X_{t-1})$
sun	*sun*	*0.9*
sun	rain	0.1
rain	*sun*	*0.3*
rain	rain	0.7

[Slide created by Dan Klein and Pieter Abbeel for CS188 Intro to AI at UC Berkeley].

Example Markov Chain Weather



- What is the probability distribution on day t ($P(X_t)$)?
- Sum out X_{t-1}

$$P(x_1) = \text{known}$$

$$P(x_t) = \sum_{x_{t-1}} P(x_{t-1}, x_t)$$
$$= \sum_{x_{t-1}} P(x_t \mid x_{t-1}) P(x_{t-1})$$

Forward simulation
Compute $P(X_2)$ then $P(X_3)$ then $P(X_4)$ …

CPT P(Xt | Xt-1):

X_{t-1}	X_t	$P(X_t \mid X_{t-1})$
sun	*sun*	*0.9*
sun	rain	0.1
rain	*sun*	*0.3*
rain	rain	0.7

[Slide created by Dan Klein and Pieter Abbeel for CS188 Intro to AI at UC Berkeley].

Example Run of Forward Computation

- From initial observation of sun

$$\underbrace{\left\langle \begin{matrix} 1.0 \\ 0.0 \end{matrix} \right\rangle}_{P(X_1)} \quad \underbrace{\left\langle \begin{matrix} 0.9 \\ 0.1 \end{matrix} \right\rangle}_{P(X_2)} \quad \underbrace{\left\langle \begin{matrix} 0.84 \\ 0.16 \end{matrix} \right\rangle}_{P(X_3)} \quad \underbrace{\left\langle \begin{matrix} 0.804 \\ 0.196 \end{matrix} \right\rangle}_{P(X_4)} \Rightarrow \underbrace{\left\langle \begin{matrix} 0.75 \\ 0.25 \end{matrix} \right\rangle}_{P(X_\infty)}$$

- From initial observation of rain

$$\underbrace{\left\langle \begin{matrix} 0.0 \\ 1.0 \end{matrix} \right\rangle}_{P(X_1)} \quad \underbrace{\left\langle \begin{matrix} 0.3 \\ 0.7 \end{matrix} \right\rangle}_{P(X_2)} \quad \underbrace{\left\langle \begin{matrix} 0.48 \\ 0.52 \end{matrix} \right\rangle}_{P(X_3)} \quad \underbrace{\left\langle \begin{matrix} 0.588 \\ 0.412 \end{matrix} \right\rangle}_{P(X_4)} \Rightarrow \underbrace{\left\langle \begin{matrix} 0.75 \\ 0.25 \end{matrix} \right\rangle}_{P(X_\infty)}$$

- From yet another initial distribution $Pr(X_1)$:

$$\underbrace{\left\langle \begin{matrix} p \\ 1-p \end{matrix} \right\rangle}_{P(X_1)} \quad \ldots \quad \Rightarrow \underbrace{\left\langle \begin{matrix} 0.75 \\ 0.25 \end{matrix} \right\rangle}_{P(X_\infty)}$$

[Slide created by Dan Klein and Pieter Abbeel for CS188 Intro to AI at UC Berkeley.]

Stationary Distributions

- For most Markov chains:
  - Influence of the initial distribution dissipates over time.
  - The distribution we end up in is independent of the initial distribution

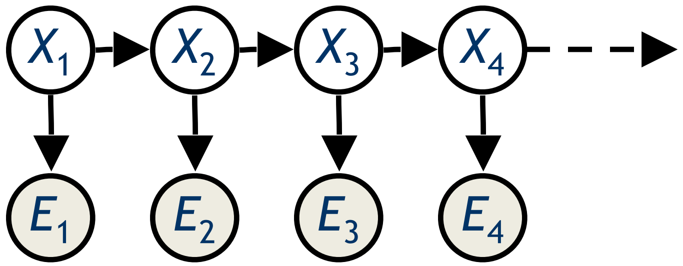
- Stationary distribution
  - The distribution that we end up with is called the stationary distribution of the chain.
  - This satisfies:

$$P_\infty(X) = P_{\infty+1}(X) = \sum_x P(X|x)P_\infty(x)$$

- That is the stationary distribution does not change on a forward progression
- We can compute it by solving simultaneous equations (or by forward simulating the system many times; forward simulation is generally computationally more effective)

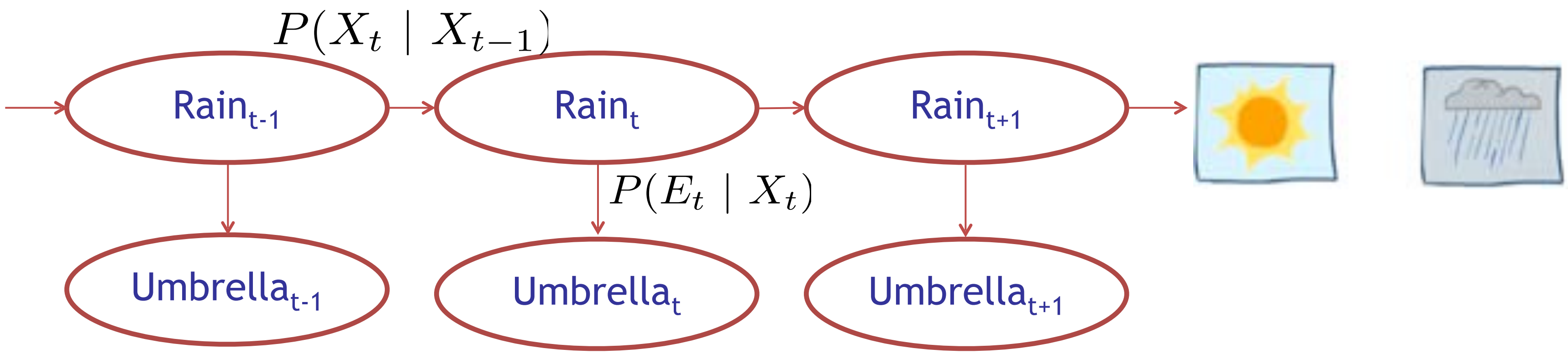
Hidden Markov Models

- Markov chains not so useful for most agents
  - Need observations to update your beliefs
- Hidden Markov models (HMMs)
  - Underlying Markov chain over states X
  - But you also observe outputs (effects) at each time step



[Slide created by Dan Klein and Pieter Abbeel for CS188 Intro to AI at UC Berkeley].

Example: Weather HMM



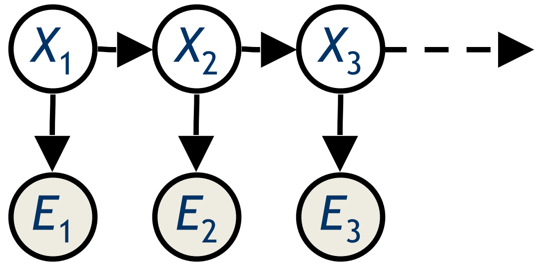
- An HMM is defined by:
  - Initial distribution: $P(X_1)$
  - Transitions: $P(X_t|X_{t-1})$
  - Emissions: $P(E_t|X_t)$

R_t	R_{t+1}	$P(R_{t+1}\|R_t)$
+r	+r	0.7
+r	-r	0.3
-r	+r	0.3
-r	-r	0.7

R_t	U_t	$P(U_t\|R_t)$
+r	+u	0.9
+r	-u	0.1
-r	+u	0.2
-r	-u	0.8

[Slide created by Dan Klein and Pieter Abbeel for CS188 Intro to AI at UC Berkeley].

Joint Distribution of an HMM



- Assumptions:
  - $P(X_t \mid X_{t-1} \ldots X_1, E_{t-1} \ldots E_1) = P(X_t \mid X_{t-1})$

    Current state is conditionally independent of early states + evidence given previous state
  - $P(X_t \mid X_{t-1})$ is the same for all time points t

    Probabilities are stationary
  - $P(E_t \mid X_t \ldots X_1, E_{t-1} \ldots E_1) = P(E_t \mid X_t)$

    Current evidence is conditionally independent of early states + early evidence given current state

Note that two evidence items are not independent, unless one of the intermediate states is known.

Real HMM Examples

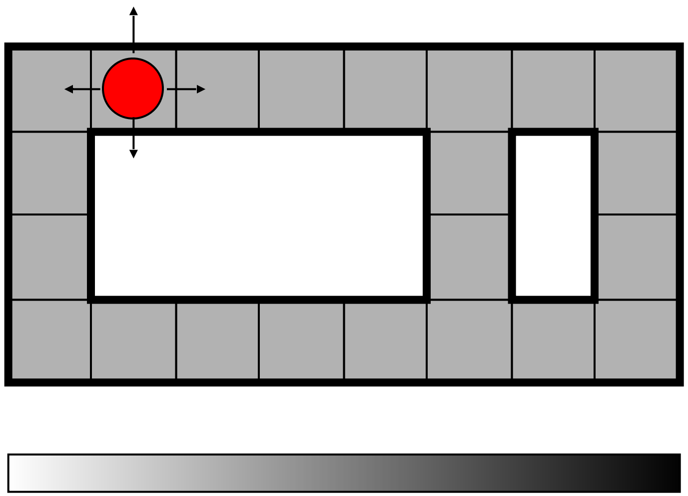
- Speech recognition HMMs:
  - Observations are acoustic signals (continuous valued)
  - States are specific positions in specific words (so, tens of thousands)
- Machine translation HMMs:
  - Observations are words (tens of thousands)
  - States are translation options
- Robot tracking:
  - Observations are range readings (continuous)
  - States are positions on a map (continuous)

[Slide created by Dan Klein and Pieter Abbeel for CS188 Intro to AI at UC Berkeley.]

Tracking/Monitoring

- Monitoring is the task of tracking $P(X_t \mid e_{t} \ldots e_1)$ over time. i.e. determining state given current and previous observations.
- $P(X_1)$ is the initial distribution over variable (or feature) X. Usually start with a uniform distribution over all values of X.
- As time elapses and we make observations and must update our distribution over X, i.e. move from $P(X_{t-1} \mid e_{t-1} \ldots e_1)$ to $P(X_t \mid e_t \ldots e_1)$.
- This means updating HMM equations. Tools to do this existed before Bayes Nets, but we can relate inference tools to Variable Elimination.

Example: Robot Localization



- Prob
- 0
- 1

Example from Michael Pfeiffer

t=0

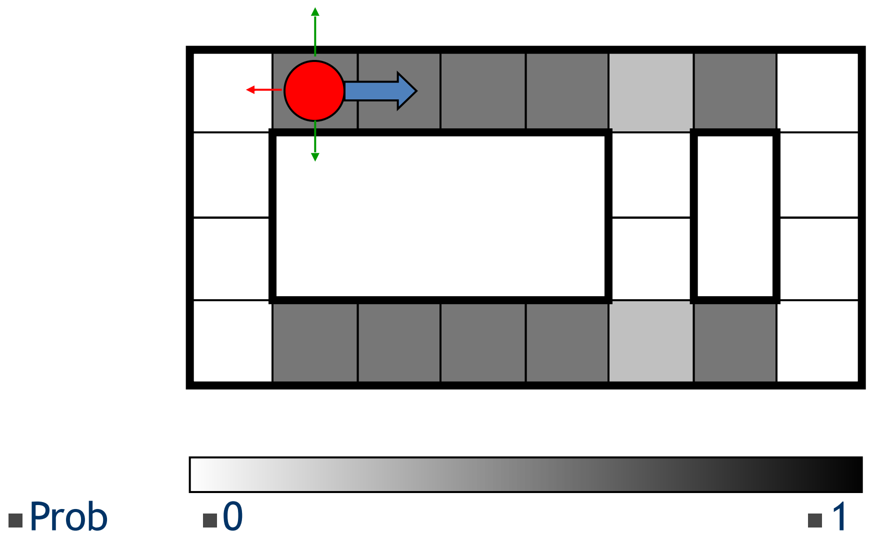
Sensor model: Can read in which directions there is a wall, never more than 1 mistake

Motion model: Either executes the move, or the robot with low probability does not move at all. Cannot move in wrong direction.

Initially uniform distribution over where robot is located—equally likely to be anywhere.

[Slide created by Dan Klein and Pieter Abbeel for CS188 Intro to AI at UC Berkeley.]

Example: Robot Localization



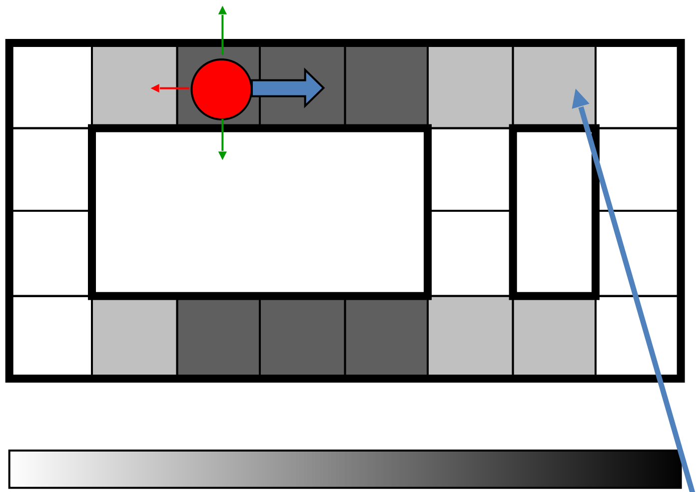
Initially don't know where you are. Observe a wall above and below, no wall to the left or right. Low probability of 1 mistake, 2 mistakes not possible

White: impossible to get this reading (more than one mistake)

Lighter grey: was possible to get the reading, but less likely because it required 1 mistake

[Slide created by Dan Klein and Pieter Abbeel for CS188 Intro to AI at UC Berkeley.]

Example: Robot Localization



- Prob
- 0
- 1

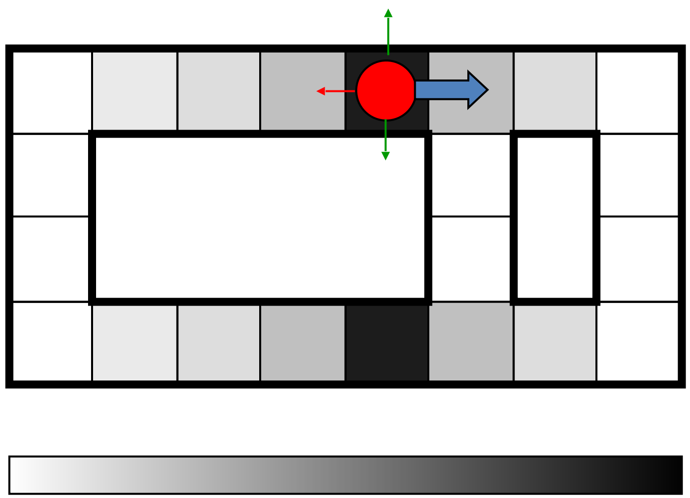
t=2: Move right. Low probability didn't move, else must have moved right.

Still observing wall above and below

can only be here if
(a) was at low probability square to the left
(b) was at this square and action didn't work.

[Slide created by Dan Klein and Pieter Abbeel for CS188 Intro to AI at UC Berkeley.]

Example: Robot Localization

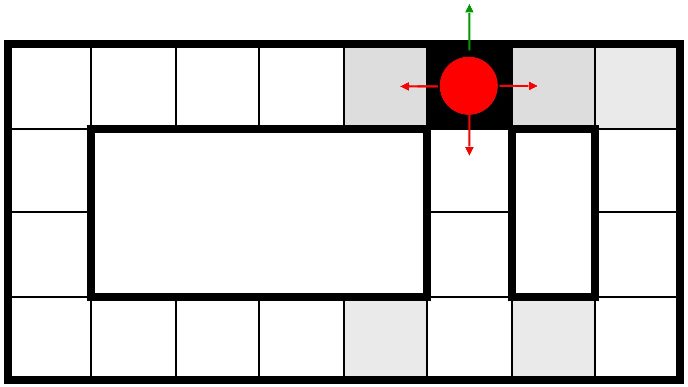


- Prob
- 0
- 1

t=4

[Slide created by Dan Klein and Pieter Abbeel for CS188 Intro to AI at UC Berkeley.]

Example: Robot Localization

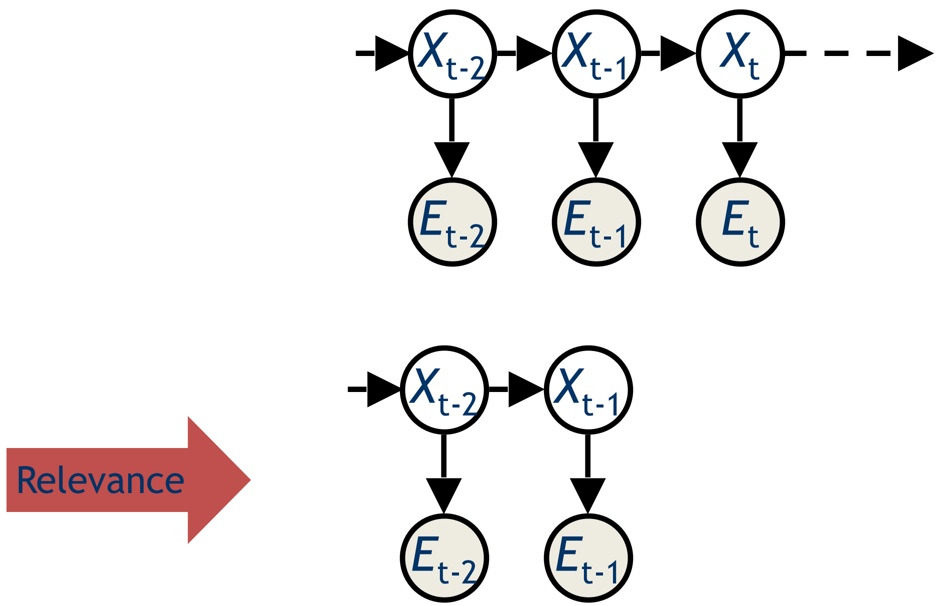


t=5

[Slide created by Dan Klein and Pieter Abbeel for CS188 Intro to AI at UC Berkeley.]

VE for $\Pr(X_{t-1}|e_{t-1},\cdots,e_1)$

- Relevance (d-separation) indicates that if X_{t-1} is the query variable, the only relevant variables are ancestors of X_{t-1}



VE for $\Pr(X_{t-1} | e_{t-1}, \cdots, e_1)$

We want

$P(X_{t-1} | e_{t-1}, e_{t-2} \ldots e_1) = P(X_{t-1}, e_{t-1}, e_{t-2} \ldots e_1) / P(e_{t-1}, e_{t-2} \ldots e_1)$.

Use VE with elimination order: $X_1, X_2 \ldots X_{t-1}$

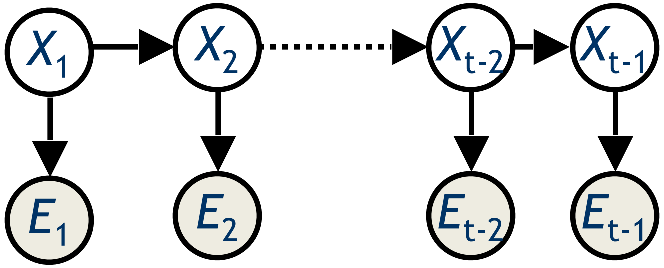
X_1: $P(X_1)\, P(e_1 | X_1) P(X_2 | X_1)$

X_2: $P(e_2 | X_2) P(X_3 | X_2)$

…

X_{t-2}: $P(e_{t-2} | X_{t-2}) P(X_{t-1} | X_{t-2})$

X_{t-1}: $P(e_{t-1} | X_{t-1})$



VE for $\Pr(X_{t-1}|e_{t-1},\cdots,e_1)$

Summing out X_1 we get a factor of X_2; summing out X_2 we get a factor of X_3 and so on:

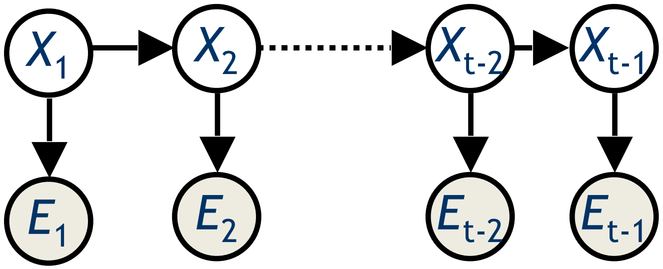
X_1: $P(X_1)\ P(e_1 | X_1)P(X_2 | X_1)$

X_2: $P(e_2 | X_2)P(X_3 | X_2)F_2(X_2)$

…

X_{t-2}: $P(e_{t-2} | X_{t-2})P(X_{t-1} | X_{t-2})F_{t-2}(X_{t-2})$

X_{t-1}: $P(e_{t-1} | X_{t-1})F_{t-1}(X_{t-1})$



VE for $\Pr(X_{t-1} | e_{t-1}, \cdots, e_1)$

X_1: $P(X_1)\ P(e_1 | X_1)P(X_2 | X_1)$

X_2: $P(e_2 | X_2)P(X_3 | X_2)F_2(X_2)$

…

X_{t-2}: $P(e_{t-2} | X_{t-2})P(X_{t-1} | X_{t-2})F_{t-2}(X_{t-2})$

X_{t-1}: $P(e_{t-1} | X_{t-1})F_{t-1}(X_{t-1})$

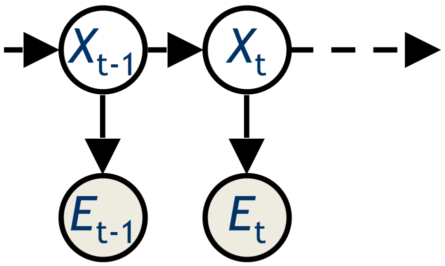
So:

$P(X_{t-1} | e_{t-1}, e_{t-2} \ldots, e_1) = \text{normalize}(P(e_{t-1} | X_{t-1})F_{t-1}(X_{t-1}))$

This is a table with one value for each X_{t-1}

VE for $\Pr(X_t | e_{t-1}, \cdots, e_1)$

Now say time has passed but no observation has been made yet.



X_1: $P(X_1)\ P(e_1 | X_1)P(X_2 | X_1)$

X_2: $P(e_2 | X_2)P(X_3 | X_2)$

…

X_{t-2}: $P(e_{t-2} | X_{t-2})P(X_{t-1} | X_{t-2})$

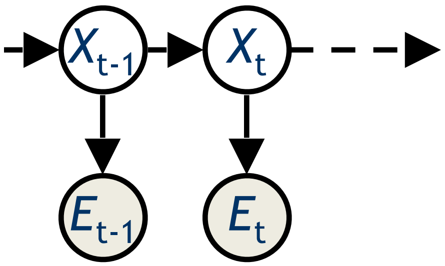
X_{t-1}: $P(e_{t-1} | X_{t-1})P(X_t | X_{t-1})$

X_t:

Same buckets with one new one (X_t) and one new factor ($P(X_t | X_{t-1})$).

VE for $\Pr(X_t | e_{t-1}, \cdots, e_1)$

Sum out variables, as before:



X_1: $P(X_1)\ P(e_1 | X_1)P(X_2 | X_1)$

X_2: $P(e_2 | X_2)P(X_3 | X_2)F_2(X_2)$

…

X_{t-2}: $P(e_{t-2} | X_{t-2})P(X_{t-1} | X_{t-2})F_{t-2}(X_{t-2})$

X_{t-1}: $P(e_{t-1} | X_{t-1})P(X_t | X_{t-1})F_{t-1}(X_{t-1})$

X_t: $F_t(X_t)$

$$F_t(X_t) = \Sigma_{d \in Dom[X_{t-1}]} P(e_{t-1} | X_{t-1})P(X_t | X_{t-1})F_{t-1}(X_{t-1})$$

VE for $\Pr(X_t | e_{t-1}, \cdots, e_1)$

We saw $P(X_{t-1} | e_{t-1}, e_{t-2} \ldots, e_1) = \text{normalize}(P(e_{t-1} | X_{t-1}) F_{t-1}(X_{t-1}))$

Means

$F_t(X_t) = \sum_{d \in Dom[X_{t-1}]} P(e_{t-1} | X_{t-1}) P(X_t | X_{t-1}) F_{t-1}(X_{t-1})$

or

$F_t(X_t) = c^* \sum_{d \in Dom[X_{t-1}]} P(X_t | X_{t-1})\, P(X_{t-1} | e_{t-1}, e_{t-2} \ldots, e_1)$

…. where c is the normalization constant.

$P(X_t | e_{t-1}, e_{t-2} \ldots, e_1) = \text{normalize}(F_t(X_t))$

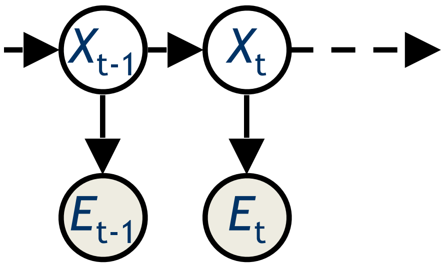
$P(X_t | e_{t-1}, e_{t-2} \ldots, e_1) =$

$\text{normalize}\left(\sum_{d \in Dom[X_{t-1}]} P(X_t | X_{t-1})\, P(X_{t-1} | e_{t-1}, e_{t-2} \ldots, e_1) \right)$

… we drop c (because we are normalizing)

VE for $\Pr(X_t | e_t, \cdots, e_1)$

How to incorporate the observation e_t? VE looks similar:



X_1: $P(X_1)\ P(e_1 | X_1) P(X_2 | X_1)$

X_2: $P(e_2 | X_2) P(X_3 | X_2) F_2(X_2)$

...

X_{t-2}: $P(e_{t-2} | X_{t-2}) P(X_{t-1} | X_{t-2}) F_{t-2}(X_{t-2})$

X_{t-1}: $P(e_{t-1} | X_{t-1}) P(X_t | X_{t-1}) F_{t-1}(X_{t-1})$

X_t: $F_t(X_t) P(e_t | X_t)$

We add $P(e_t | X_t)$ to the bucket for X_t and normalize.

VE for $\Pr(X_t | e_t, \cdots, e_1)$

So $P(X_t | e_t, e_{t-1}, \ldots e_1) = F_t(X_t) P(e_t | X_t)$

We saw that

$P(X_t | e_{t-1}, e_{t-2} \ldots, e_1) = \text{normalize}(F_t(X_t)) = c*F_t(X_t)$

So

$P(X_t | e_t, e_{t-1}, e_{t-2} \ldots, e_1) = \text{normalize}(c*F_t(X_t)*P(e_t | X_t))$

$= \text{normalize}(F_t(X_t)*P(e_t | X_t))$

… we again drop c (because we are normalizing)

HMM Rules, Recap

1. Access initial distribution ($P(X_1)$)
2. Calculate state estimates over time:

$P(X_t | e_{t-1}, e_{t-2} \ldots, e_1) =$

normalize($\sum_{d \in Dom[X_{t-1}]} P(X_t | X_{t-1}) P(X_{t-1} | e_{t-1}, e_{t-2} \ldots, e_1)$)

3. Weight with observation:

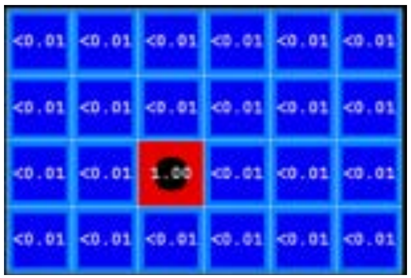
$P(X_t | e_t, e_{t-1}, e_{t-2} \ldots, e_1) =$

normalize($P(X_t | e_{t-1}, e_{t-2} \ldots, e_1) * P(e_t | X_t)$)

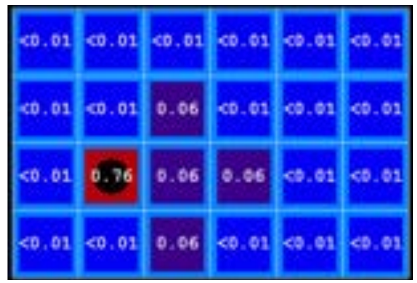
Example: Passage of Time

- As time passes, uncertainty "accumulates"

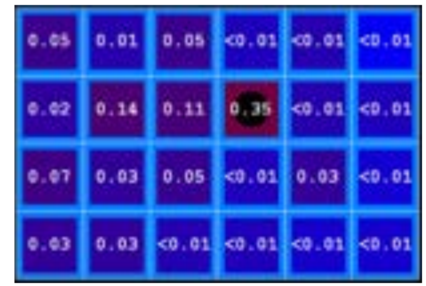
■ (Transition model: ghosts usually go clockwise)



T = 1



T = 2

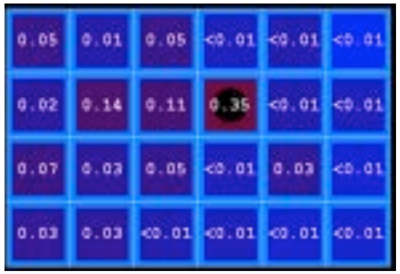


T = 5

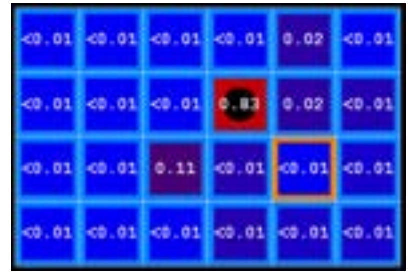
[Slide created by Dan Klein and Pieter Abbeel for CS188 Intro to AI at UC Berkeley.]

Example: Observation

- As we get observations, beliefs get re-weighted, uncertainty "decreases"



Before observation



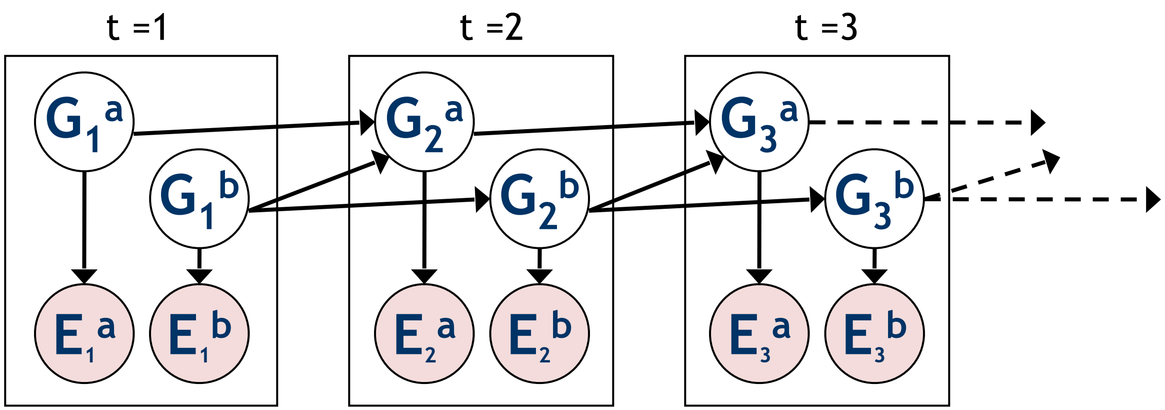
After observation

$$P(X_t \mid e_t, e_{t-1} \dots, e_1) = c * (P(X_t \mid e_{t-1}, e_{t-2} \dots, e_1) * P(e_t \mid X_t))$$

[Slide created by Dan Klein and Pieter Abbeel for CS188 Intro to AI at UC Berkeley.]

Dynamic Bayes Nets (DBNs)

- Track multiple variables over time, using multiple sources of evidence
- Idea: repeat a fixed Bayes net structure at each time
- Variables from time *t* can be conditional on those from *t-1*



- *Dynamic Bayes nets are a generalization of HMMs*

[Slide created by Dan Klein and Pieter Abbeel for CS188 Intro to AI at UC Berkeley.]