复旦大学大数据学院
School of Data Science, Fudan University
魏忠钰

Markov Models and Hidden Markov Models

May 16th, 2017

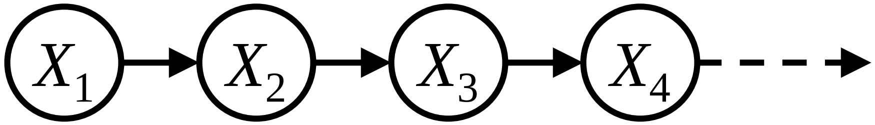
Reasoning over Time or Space

- Often, 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

- Value of X at a given time is called the state



- Parameters: called transition probabilities or dynamics, specify how the state evolves over time (also, initial state probabilities)
- Same as MDP transition model, but no choice of action

Chain Rule and Markov Models

$$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$$

- From the chain rule, every joint distribution over X_1, X_2, X_3, X_4

can be written as:

$$P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)P(X_4|X_1, X_2, X_3)$$

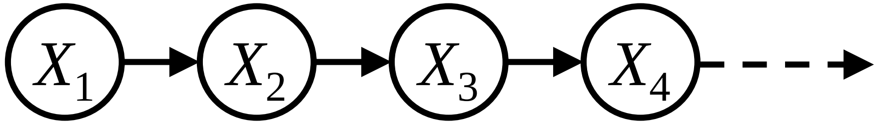
- Assuming that

$$X_3 \perp\!\!\!\perp X_1 \mid X_2 \qquad \text{and} \qquad X_4 \perp\!\!\!\perp X_1, X_2 \mid X_3$$

results in the expression posited on the previous slide:

$$P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3|X_2)P(X_4|X_3)$$

Chain Rule and Markov Models



- From the chain rule, every joint distribution over $X_1, X_2, \ldots, X_T$ can be written as:

$$P(X_1, X_2, \ldots, X_T) = P(X_1) \prod_{t=2}^{T} P(X_t | X_1, X_2, \ldots, X_{t-1})$$

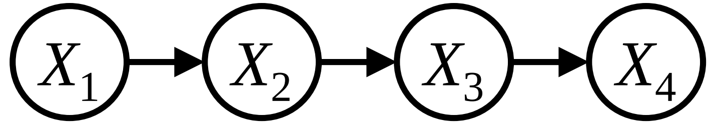
- Assuming that for all *t*:

$$X_t \perp\!\!\!\perp X_1, \ldots, X_{t-2} \mid X_{t-1}$$

gives us the expression posited on the earlier slide:

$$P(X_1, X_2, \ldots, X_T) = P(X_1) \prod_{t=2}^{T} P(X_t | X_{t-1})$$

Joint Distribution of a Markov Model



$P(X_1) \qquad P(X_t|X_{t-1})$

- Joint distribution:

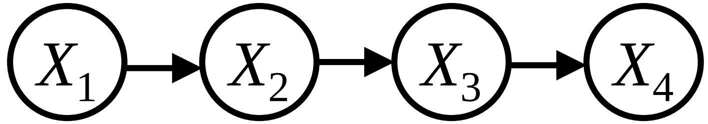
$$P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3|X_2)P(X_4|X_3)$$

- More generally:

$$\begin{aligned} P(X_1, X_2, \ldots, X_T) &= P(X_1)P(X_2|X_1)P(X_3|X_2)\ldots P(X_T|X_{T-1}) \\ &= P(X_1)\prod_{t=2}^{T} P(X_t|X_{t-1}) \end{aligned}$$

- Stationarity assumption: transition probabilities the same at all times
- Other conditional independency?

Implied Conditional Independencies

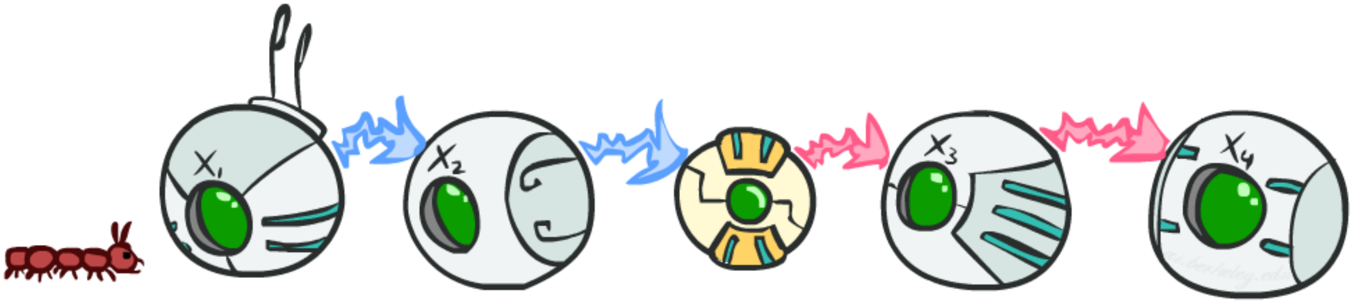


- We assumed: $X_3 \perp\!\!\!\perp X_1 \mid X_2$ and $X_4 \perp\!\!\!\perp X_1, X_2 \mid X_3$

- Do we also have $X_1 \perp\!\!\!\perp X_3, X_4 \mid X_2$?
  - Yes!
  - Proof:

$$
\begin{aligned}
P(X_1 \mid X_2, X_3, X_4) &= \frac{P(X_1, X_2, X_3, X_4)}{P(X_2, X_3, X_4)} \\
&= \frac{P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_2)P(X_4 \mid X_3)}{\sum_{x_1} P(x_1)P(X_2 \mid x_1)P(X_3 \mid X_2)P(X_4 \mid X_3)} \\
&= \frac{P(X_1, X_2)}{P(X_2)} \\
&= P(X_1 \mid X_2)
\end{aligned}
$$

Conditional Independence



- Basic conditional independence:
  - Past and future independent of the present
  - Each time step only depends on the previous
  - This is called the (first order) Markov property

Markov Models Recap

- Explicit assumption for all *t* : $X_t \perp\!\!\!\perp X_1, \ldots, X_{t-2} \mid X_{t-1}$
- Consequence, joint distribution can be written as:

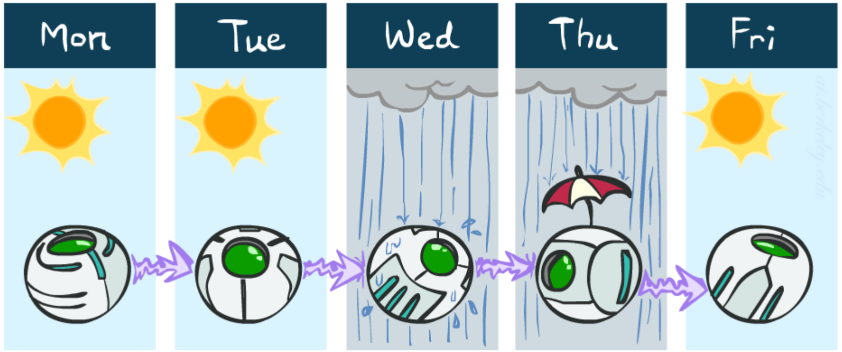
$$P(X_1, X_2, \ldots, X_T) = P(X_1)P(X_2|X_1)P(X_3|X_2)\ldots P(X_T|X_{T-1})$$
$$= P(X_1)\prod_{t=2}^{T} P(X_t|X_{t-1})$$

- Implied conditional independencies:
  - Past variables independent of future variables given the present

i.e., if $t_1 < t_2 < t_3$ or $t_1 > t_2 > t_3$ then: $X_{t_1} \perp\!\!\!\perp X_{t_3} \mid X_{t_2}$

- Additional explicit assumption:
  - $P(X_t \mid X_{t-1})$ is the same for all *t*

Example Markov Chain: Weather



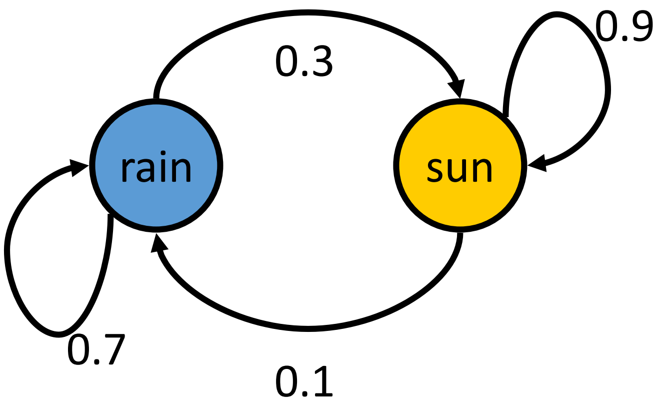
- States: X = {rain, sun}

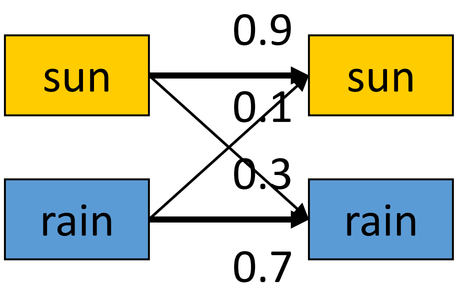
- Initial distribution: 1.0 sun

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

Two new ways of representing the same CPT

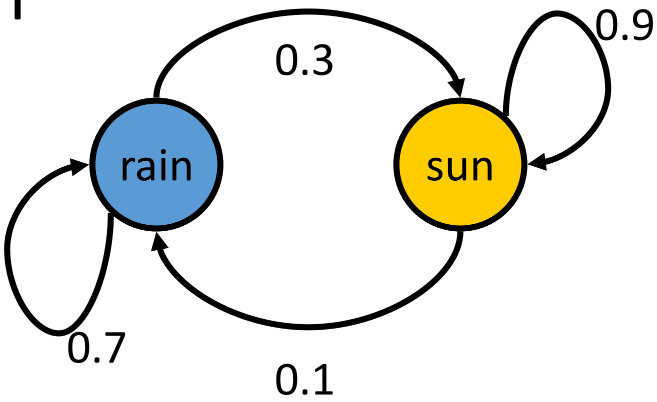
X_{t-1}	X_t	$P(X_t \mid X_{t-1})$
sun	sun	0.9
sun	rain	0.1
rain	sun	0.3
rain	rain	0.7





Example Markov Chain: Weather

- Initial distribution: 1.0 sun



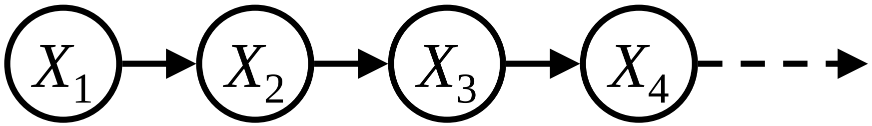
- What is the probability distribution after one step?

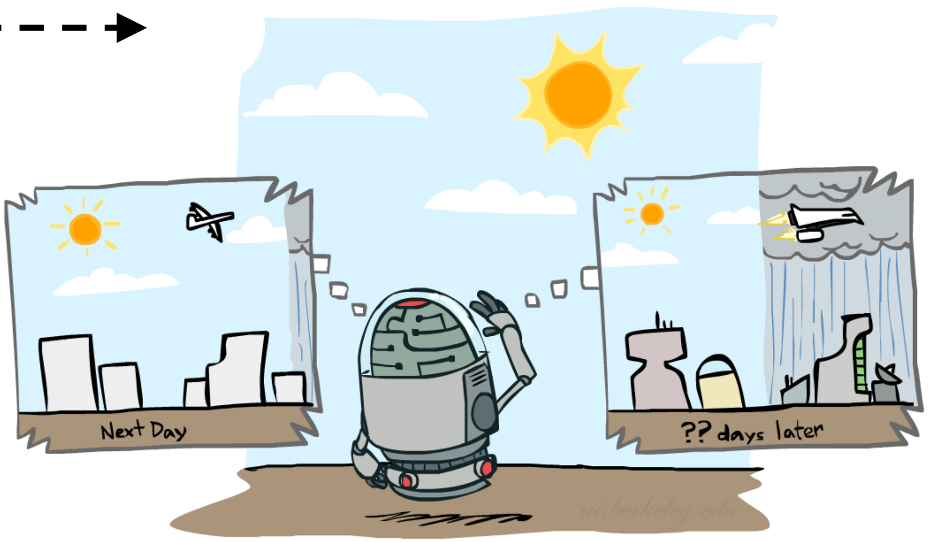
$$P(X_2 = \text{sun}) = \quad P(X_2 = \text{sun} | X_1 = \text{sun}) P(X_1 = \text{sun}) + P(X_2 = \text{sun} | X_1 = \text{rain}) P(X_1 = \text{rain})$$

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

Mini-Forward Algorithm

- Question: What's P(X) on some day t?





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

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

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

Forward simulation

Example Run of Mini-Forward Algorithm

- From initial observation of sun

$$\left\langle \begin{matrix} 1.0 \\ 0.0 \end{matrix} \right\rangle \quad \left\langle \begin{matrix} 0.9 \\ 0.1 \end{matrix} \right\rangle \quad \left\langle \begin{matrix} 0.84 \\ 0.16 \end{matrix} \right\rangle \quad \left\langle \begin{matrix} 0.804 \\ 0.196 \end{matrix} \right\rangle \Rightarrow \left\langle \begin{matrix} 0.75 \\ 0.25 \end{matrix} \right\rangle$$

$$\mathrm{P}(X_1) \quad \mathrm{P}(X_2) \quad \mathrm{P}(X_3) \quad \mathrm{P}(X_4) \quad \mathrm{P}(X_\infty)$$

- From initial observation of rain

$$\left\langle \begin{matrix} 0.0 \\ 1.0 \end{matrix} \right\rangle \quad \left\langle \begin{matrix} 0.3 \\ 0.7 \end{matrix} \right\rangle \quad \left\langle \begin{matrix} 0.48 \\ 0.52 \end{matrix} \right\rangle \quad \left\langle \begin{matrix} 0.588 \\ 0.412 \end{matrix} \right\rangle \Rightarrow \left\langle \begin{matrix} 0.75 \\ 0.25 \end{matrix} \right\rangle$$

$$\mathrm{P}(X_1) \quad \mathrm{P}(X_2) \quad \mathrm{P}(X_3) \quad \mathrm{P}(X_4) \quad \mathrm{P}(X_\infty)$$

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

$$\left\langle \begin{matrix} p \\ 1-p \end{matrix} \right\rangle \quad \cdots \quad \Rightarrow \left\langle \begin{matrix} 0.75 \\ 0.25 \end{matrix} \right\rangle$$

$$\mathrm{P}(X_1) \qquad \mathrm{P}(X_\infty)$$

Stationary Distributions

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

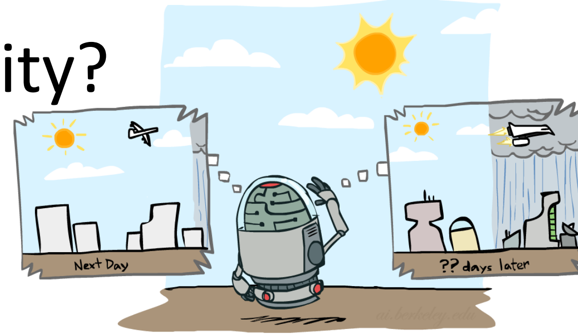
- Stationary distribution:
  - The distribution we end up with is called the stationary distribution P_∞ of the chain
  - It satisfies

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

Example: Stationary Distributions

- Question: What's P(X) at time t = infinity?

$$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4 \dashrightarrow$$



$$P_\infty(sun) = P(sun|sun)P_\infty(sun) + P(sun|rain)P_\infty(rain)$$
$$P_\infty(rain) = P(rain|sun)P_\infty(sun) + P(rain|rain)P_\infty(rain)$$

$$P_\infty(sun) = 0.9P_\infty(sun) + 0.3P_\infty(rain)$$
$$P_\infty(rain) = 0.1P_\infty(sun) + 0.7P_\infty(rain)$$

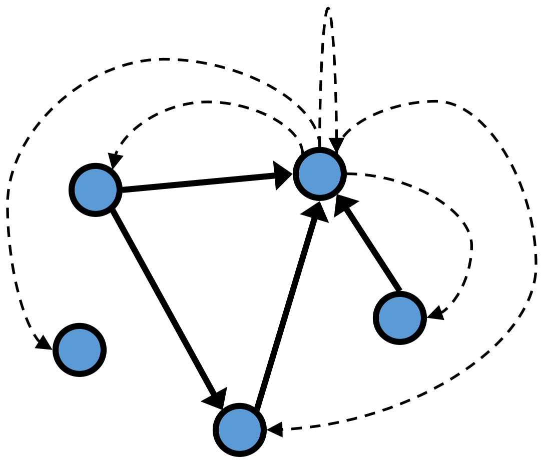
$$P_\infty(sun) = 3P_\infty(rain)$$
$$P_\infty(rain) = 1/3P_\infty(sun)$$

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

Also: $P_\infty(sun) + P_\infty(rain) = 1$

$$P_\infty(sun) = 3/4$$
$$P_\infty(rain) = 1/4$$

Stationary Distribution for Weblink analysis

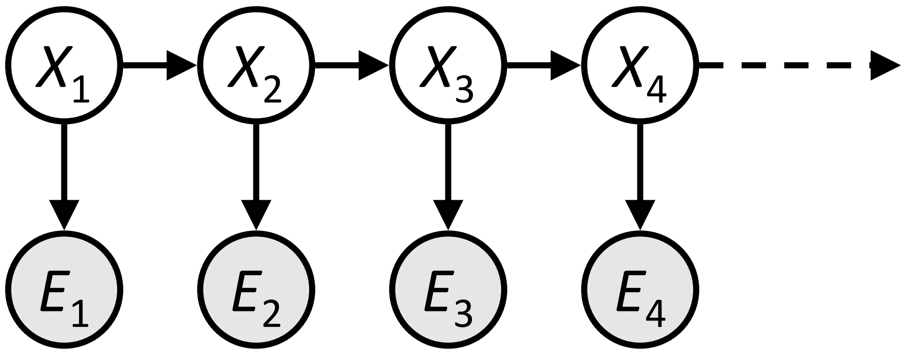


- PageRank over a web graph
  - Each web page is a state
  - Initial distribution: uniform over pages
  - Transitions:
    - With prob. c, uniform jump to a random page (dotted lines, not all shown)
    - With prob. 1-c, follow a random outlink (solid lines)
- Stationary distribution
  - Will spend more time on highly reachable pages
  - E.g. many ways to get to the Acrobat Reader download page
  - Somewhat robust to link spam
  - Google 1.0 returned the set of pages containing all your keywords in decreasing rank, now all search engines use link analysis along with many other factors (rank actually getting less important over time)

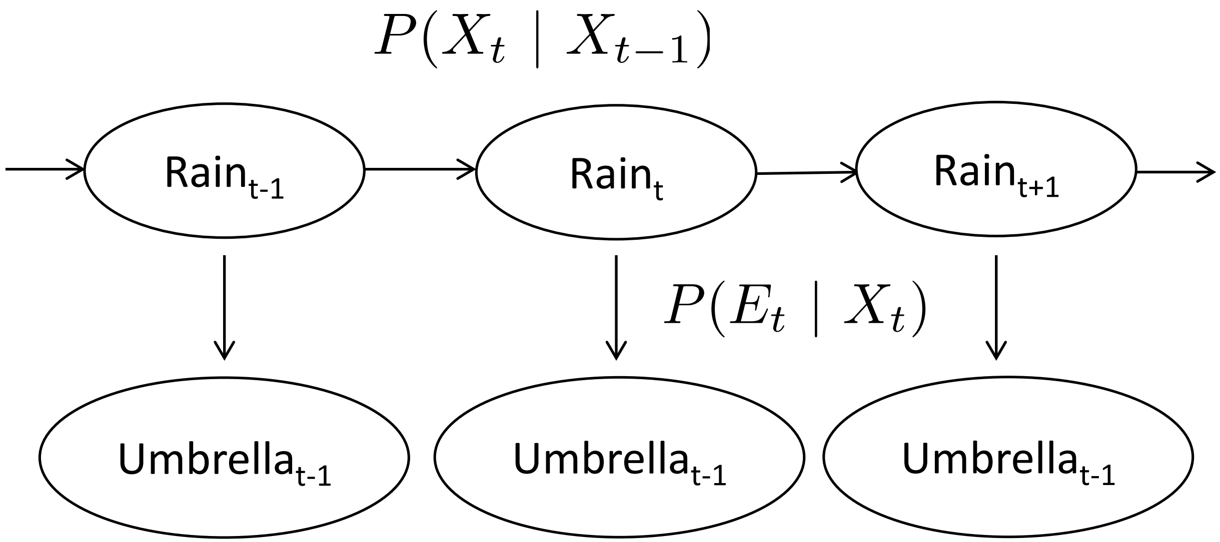
Hidden Markov Model

- Markov chains not so useful for most agents
  - Need observations to update your beliefs

- Transition model + Sensor model
- Hidden Markov models (HMMs)
  - Underlying Markov chain over states X
  - You observe outputs (effects) at each time step



Example: Weather HMM



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

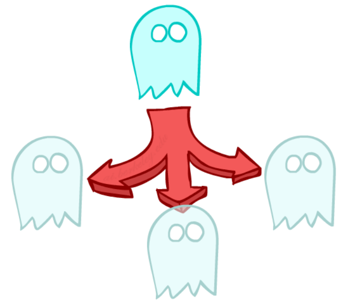
- An HMM is defined by:
  - Initial distribution: $P(X_1)$
  - Transitions: $P(X_t \mid X_{t-1})$
  - Emissions: $P(E_t \mid X_t)$

Example: Ghostbusters HMM

- $P(X_1)$ = uniform

- $P(X|X')$ = usually move clockwise, but sometimes move in a random direction or stay in place

- $P(R_{ij}|X)$ = same sensor model as before: red means close, green means far away.

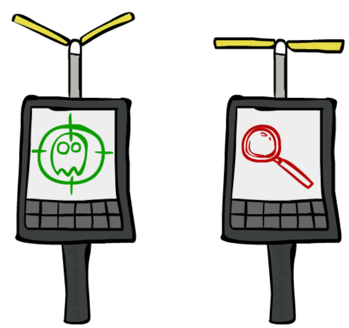


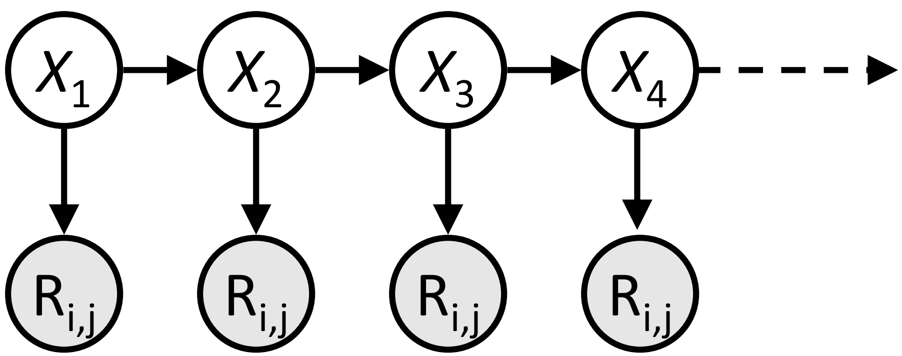
1/9	1/9	1/9
1/9	1/9	1/9
1/9	1/9	1/9

$P(X_1)$

1/6	1/6	1/2
0	1/6	0
0	0	0

$P(X|X' = <1,2>)$





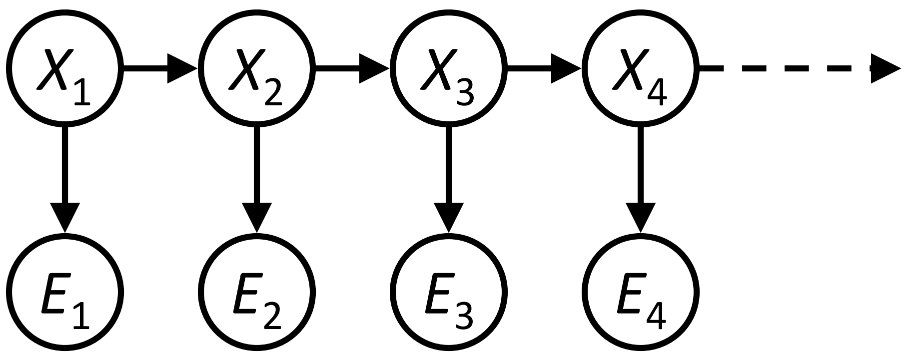
Tasks for HMM

- **Filtering**
  - Computing the **belief state**—the posterior distribution over the most recent state—given all evidence to date.
  - $\boldsymbol{P(X_t|e_{1:t})}$
- **Prediction**
  - Computing the posterior distribution over the *future* state, given all evidence to date.
  - $\boldsymbol{P(X_{t+k}|e_{1:t})}$
- **Smoothing**
  - Computing the posterior distribution over a *past* state, given all evidence up to the present.
  - $\boldsymbol{P(X_k|e_{1:t})}$
- **Most Likely Explanation**
  - Given a sequence of observations, find the sequence of states that is most likely to have generated those observations.

$$\arg\max_{x_{1:t}} P(x_{1:t}|e_{1:t})$$

Conditional Independence

- HMMs have two important independence properties:
  - Markov hidden process: future depends on past via the present
  - Current observation independent of all else given current state



- Does this mean that evidence variables are guaranteed to be independent?
  - [No, they tend to correlated by the hidden state]

Chain Rule and HMMs

- From the chain rule, *every* joint distribution over $X_1, E_1, \ldots, X_T, E_T$

can be written as:

$$P(X_1, E_1, \ldots, X_T, E_T) = P(X_1)P(E_1|X_1)\prod_{t=2}^{T} P(X_t|X_1, E_1, \ldots, X_{t-1}, E_{t-1})P(E_t|X_1, E_1, \ldots, X_{t-1}, E_{t-1}, X_t)$$

- *Assuming* that for all *t*:
  - State independent of all past states and all past evidence given the previous state, i.e.:

$$X_t \perp\!\!\!\perp X_1, E_1, \ldots, X_{t-2}, E_{t-2}, E_{t-1} \mid X_{t-1}$$

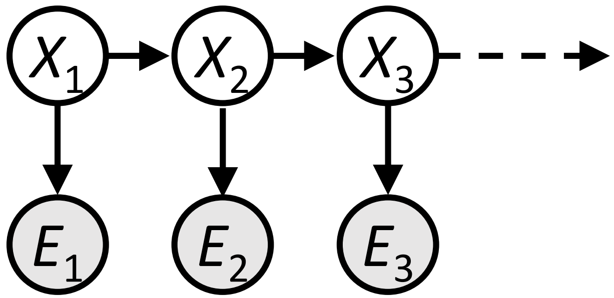
  - Evidence is independent of all past states and all past evidence given the current state, i.e.:

$$E_t \perp\!\!\!\perp X_1, E_1, \ldots, X_{t-2}, E_{t-2}, X_{t-1}, E_{t-1} \mid X_t$$

So, we have:

$$P(X_1, E_1, \ldots, X_T, E_T) = P(X_1)P(E_1|X_1)\prod_{t=2}^{T} P(X_t|X_{t-1})P(E_t|X_t)$$

Implied Conditional Independencies



- Many implied conditional independencies, e.g.,

$$E_1 \perp\!\!\!\perp X_2, E_2, X_3, E_3 \mid X_1$$

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)

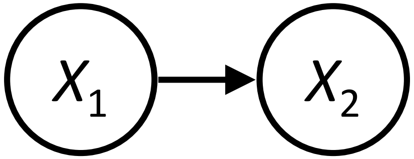
Tasks for HMM

- **Prediction**
  - Computing the posterior distribution over the *future* state, given all evidence to date.
  - $\boldsymbol{P}(\boldsymbol{X}_{\boldsymbol{t}+\boldsymbol{k}}|\boldsymbol{e}_{\mathbf{1}:\boldsymbol{t}})$

Prediction

- Assume we have current belief P(X | evidence to date)

$$B(X_t) = P(X_t|e_{1:t})$$



- Then, after one time step passes:

$$P(X_{t+1}|e_{1:t}) = \sum_{x_t} P(X_{t+1}, x_t|e_{1:t})$$

$$= \sum_{x_t} P(X_{t+1}|x_t, e_{1:t})P(x_t|e_{1:t})$$

$$= \sum_{x_t} P(X_{t+1}|x_t)P(x_t|e_{1:t})$$

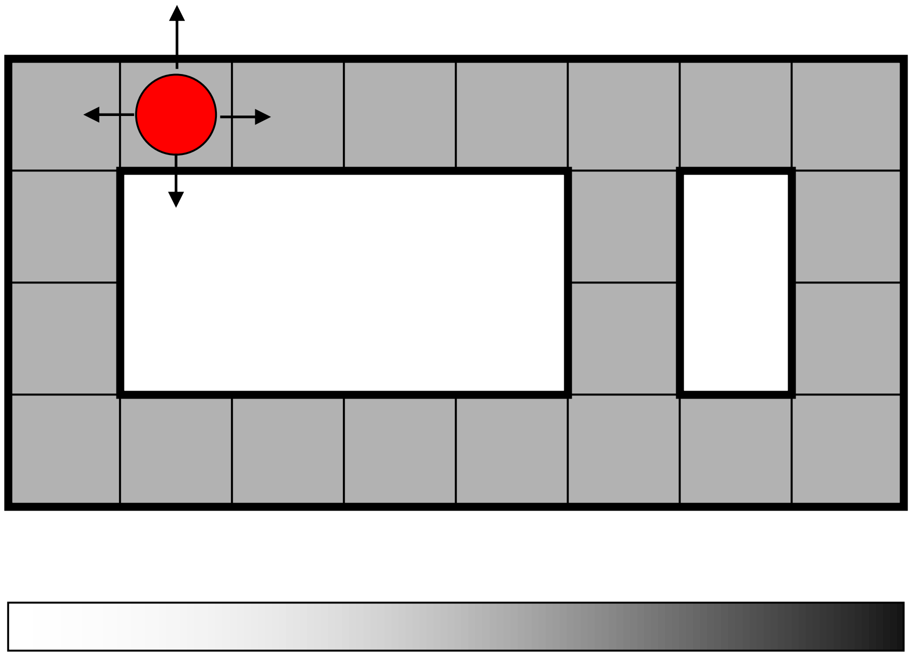
- Or compactly:

$$B'(X_{t+1}) = \sum_{x_t} P(X'|x_t)B(x_t)$$

Filtering / Monitoring

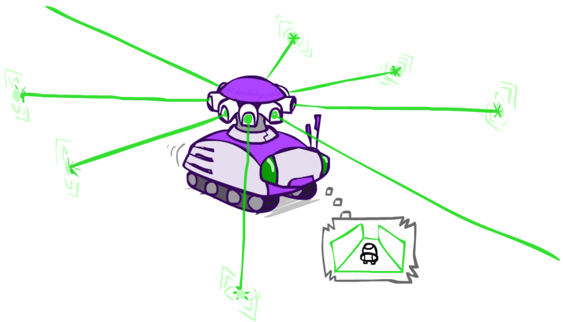
- Filtering, or monitoring, is the task of tracking the distribution $B_t(X) = P_t(X_t \mid e_1, \ldots, e_t)$ (the belief state) over time
- We start with $B_1(X)$ in an initial setting, usually uniform
- As time passes, or we get observations, we update $B(X)$
- The Kalman filter was invented in the 60's and first implemented as a method of trajectory estimation for the Apollo program

Example: Robot Localization



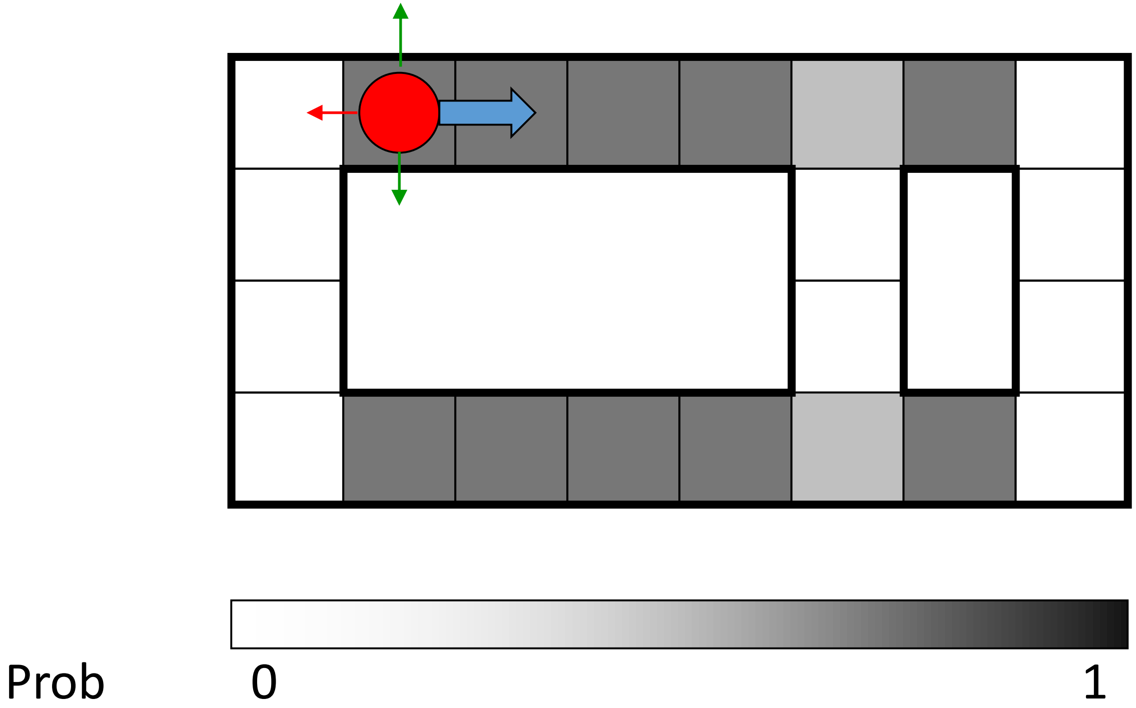
Prob 0 1

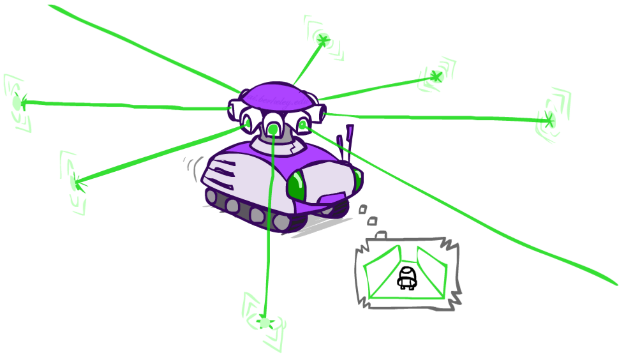
t=0



- Sensor model: can read in which directions there is a wall, never more than 1 mistake.
- Motion model: may not execute action with small prob.

Example: Robot Localization

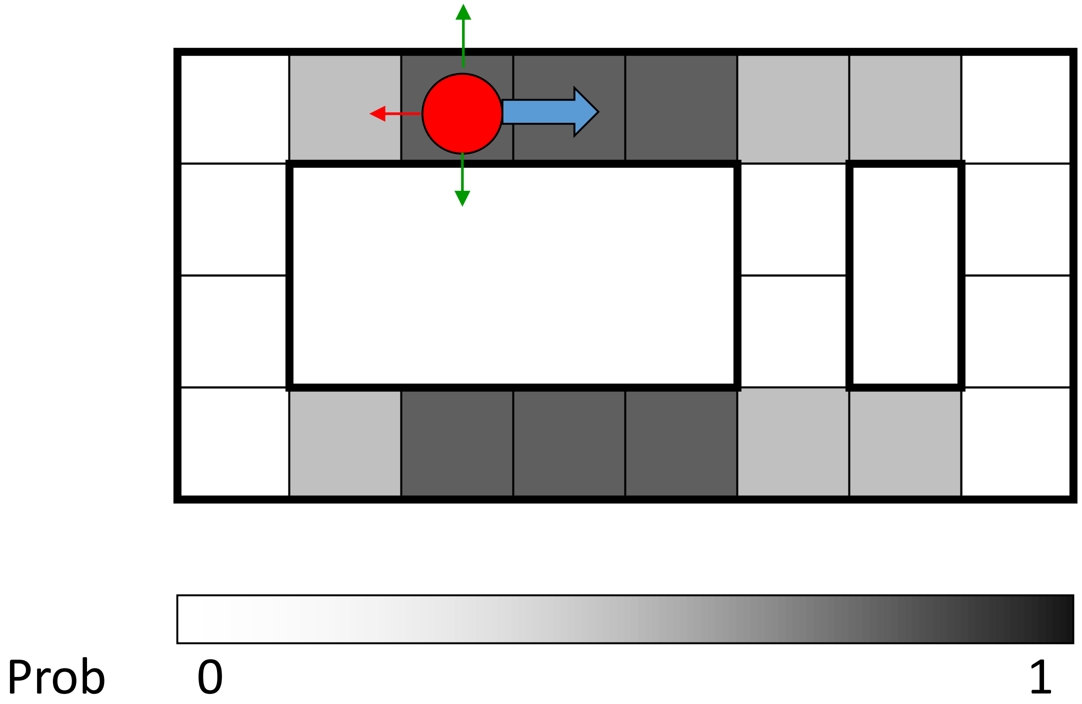




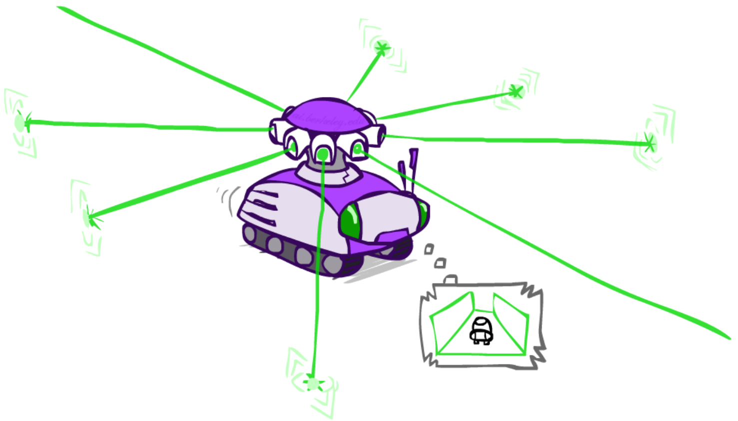
t=1

Lighter grey: was possible to get the reading, but less likely b/c required 1 mistake

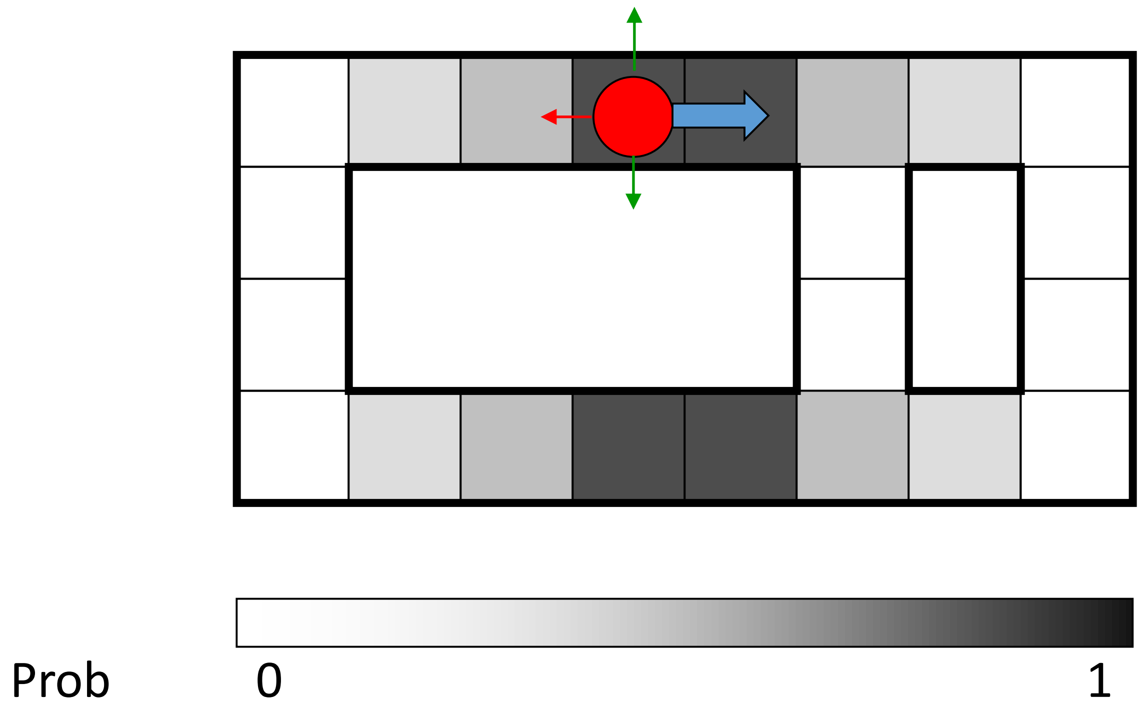
Example: Robot Localization



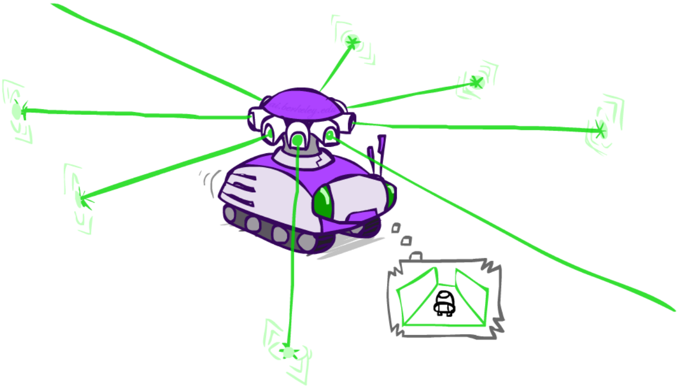
t=2



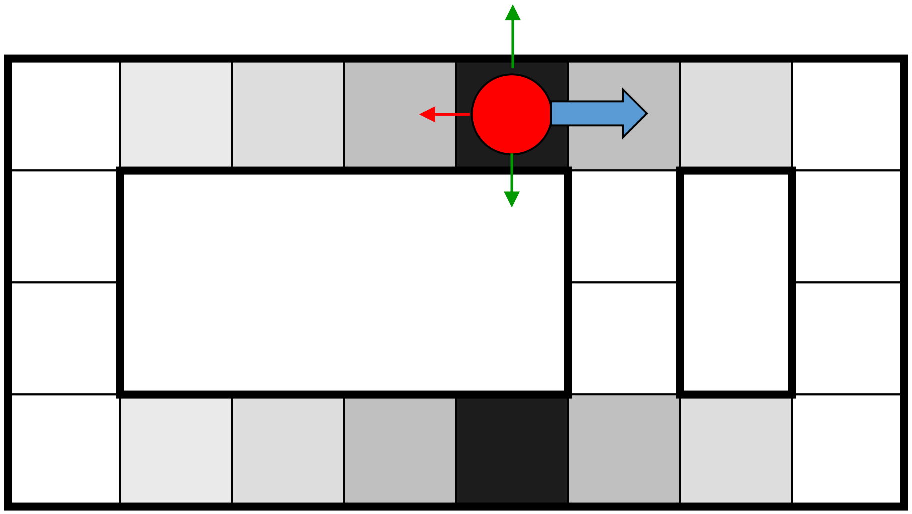
Example: Robot Localization

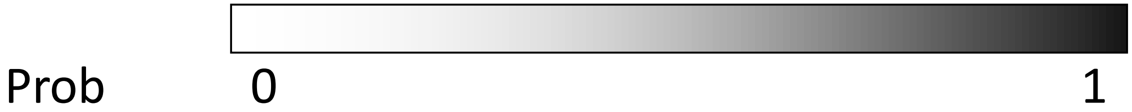


t=3

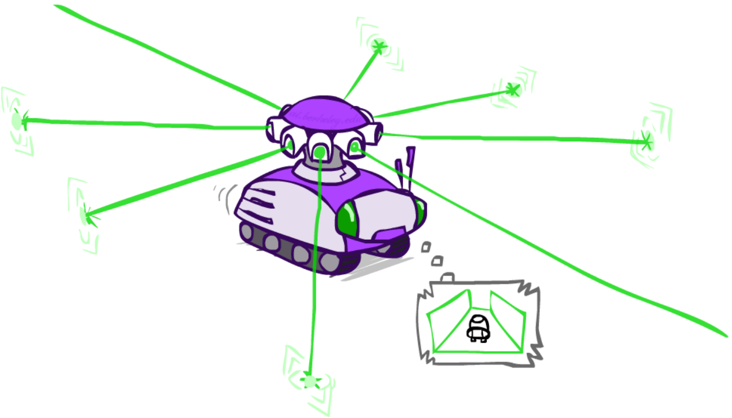


Example: Robot Localization

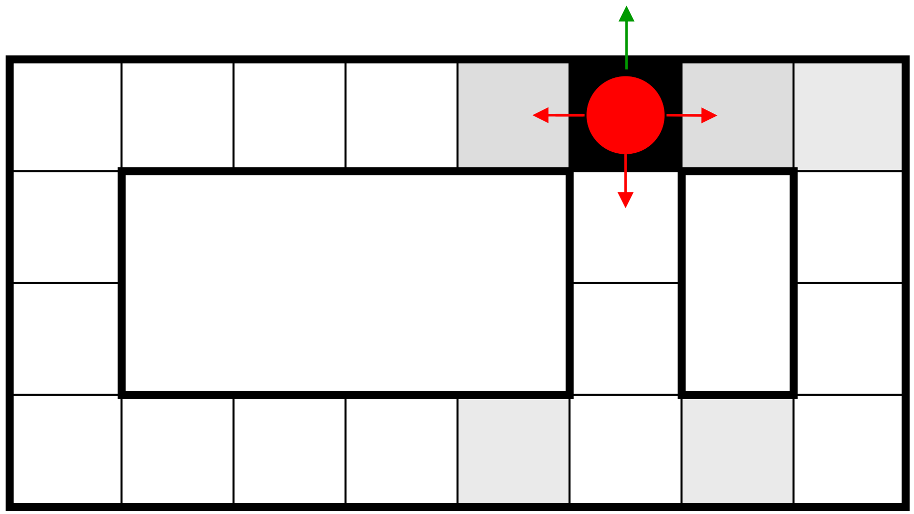




t=4

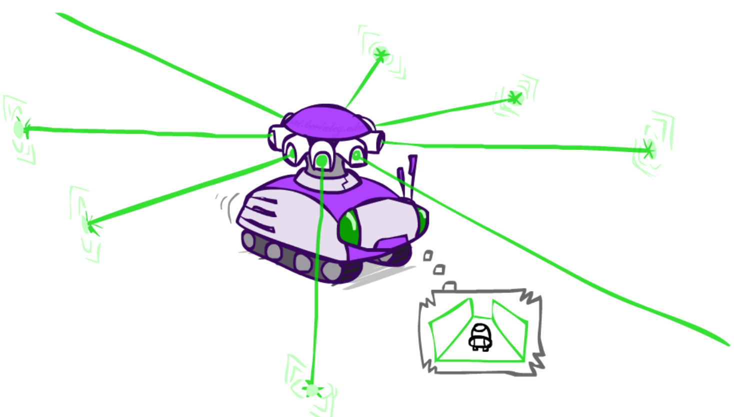


Example: Robot Localization

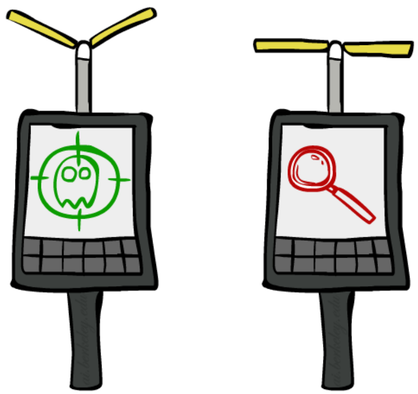


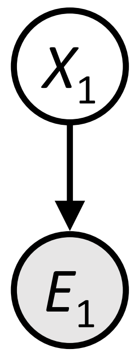
Prob 0 1

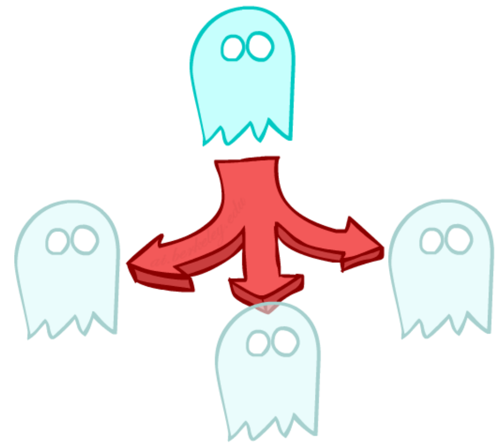
t=5

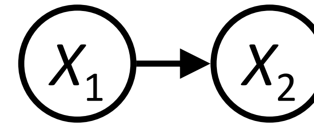


Inference: Base Cases









$$P(X_1|e_1)$$

$$\begin{aligned} P(x_1|e_1) &= P(x_1, e_1)/P(e_1) \\ &\propto_{X_1} P(x_1, e_1) \\ &= P(x_1)P(e_1|x_1) \end{aligned}$$

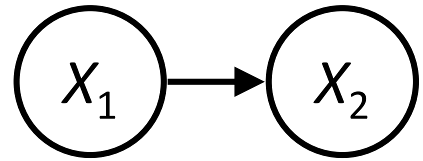
$$P(X_2)$$

$$\begin{aligned} P(x_2) &= \sum_{x_1} P(x_1, x_2) \\ &= \sum_{x_1} P(x_1)P(x_2|x_1) \end{aligned}$$

Passage of Time

- Assume we have current belief P(X | evidence to date)

$$B(X_t) = P(X_t|e_{1:t})$$



- Then, after one time step passes:

$$P(X_{t+1}|e_{1:t}) = \sum_{x_t} P(X_{t+1}, x_t|e_{1:t})$$

$$= \sum_{x_t} P(X_{t+1}|x_t, e_{1:t}) P(x_t|e_{1:t})$$

$$= \sum_{x_t} P(X_{t+1}|x_t) P(x_t|e_{1:t})$$

- Or compactly:

$$B'(X_{t+1}) = \sum_{x_t} P(X'|x_t) B(x_t)$$

Example: Passage of Time

- As time passes, uncertainty “accumulates”

<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
<0.01	<0.01	1.00	<0.01	<0.01	<0.01
<0.01	<0.01	<0.01	<0.01	<0.01	<0.01

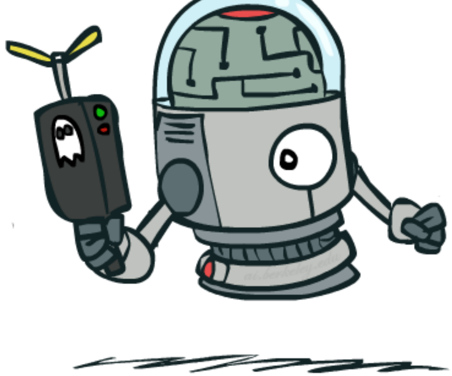
T = 1

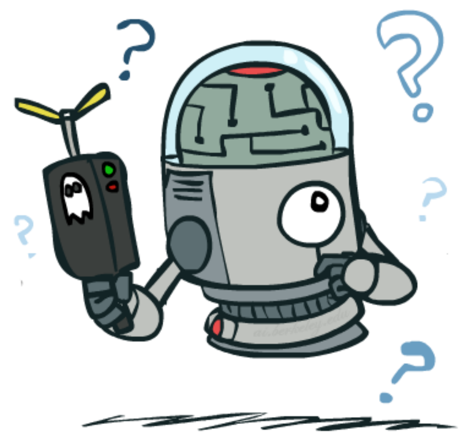
<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
<0.01	<0.01	0.06	<0.01	<0.01	<0.01
<0.01	0.76	0.06	0.06	<0.01	<0.01
<0.01	<0.01	0.06	<0.01	<0.01	<0.01

T = 2

0.05	0.01	0.05	<0.01	<0.01	<0.01
0.02	0.14	0.11	0.35	<0.01	<0.01
0.07	0.03	0.05	<0.01	0.03	<0.01
0.03	0.03	<0.01	<0.01	<0.01	<0.01

T = 5





Observation

- Assume we have current belief P(X | previous evidence):

$$B'(X_{t+1}) = P(X_{t+1}|e_{1:t})$$

- Then, after evidence comes in:

$$\begin{aligned} P(X_{t+1}|e_{1:t+1}) = \quad & P(X_{t+1}, e_{t+1}|e_{1:t})/P(e_{t+1}|e_{1:t}) \\ & \propto_{X_{t+1}} P(X_{t+1}, e_{t+1}|e_{1:t}) \\ & = P(e_{t+1}|e_{1:t}, X_{t+1})P(X_{t+1}|e_{1:t}) \\ & = P(e_{t+1}|X_{t+1})P(X_{t+1}|e_{1:t}) \end{aligned}$$

- Or, compactly:

$$B(X_{t+1}) \propto_{X_{t+1}} P(e_{t+1}|X_{t+1})B'(X_{t+1})$$

- Basic idea: beliefs "reweighted" by likelihood of evidence
- Unlike passage of time, we have to renormalize

Example: Observation

- As we get observations, beliefs get reweighted, uncertainty "decreases"

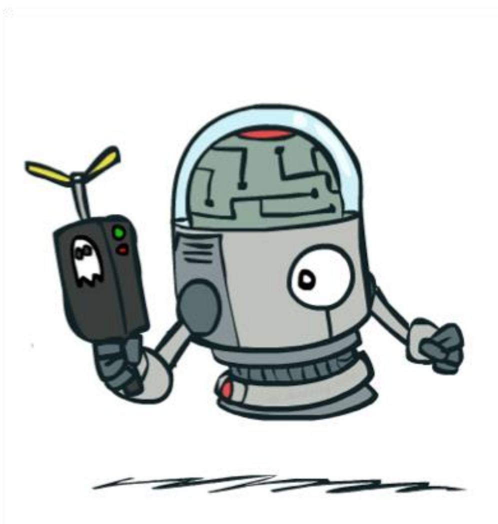
0.05	0.01	0.05	<0.01	<0.01	<0.01
0.02	0.14	0.11	0.35	<0.01	<0.01
0.07	0.03	0.05	<0.01	0.03	<0.01
0.03	0.03	<0.01	<0.01	<0.01	<0.01

Before observation

<0.01	<0.01	<0.01	<0.01	0.02	<0.01
<0.01	<0.01	<0.01	0.83	0.02	<0.01
<0.01	<0.01	0.11	<0.01	<0.01	<0.01
<0.01	<0.01	<0.01	<0.01	<0.01	<0.01

After observation

$$B(X) \propto P(e|X)B'(X)$$



The Forward Algorithm

- We are given evidence at each time and want to know

$$B_t(X) = P(X_t|e_{1:t})$$

- We can derive the following updates

We can normalize as we go if we want to have P(x|e) at each time step, or just once at the end…

$$P(x_t|e_{1:t}) \propto_X P(x_t, e_{1:t})$$

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

$$= \sum_{x_{t-1}} P(x_{t-1}, e_{1:t-1})P(x_t|x_{t-1})P(e_t|x_t)$$

$$= P(e_t|x_t) \sum_{x_{t-1}} P(x_t|x_{t-1})P(x_{t-1}, e_{1:t-1})$$

Online Belief Updates

- Every time step, we start with current P(X | evidence)
- We update for time:

$$P(x_t|e_{1:t-1}) = \sum_{x_{t-1}} P(x_{t-1}|e_{1:t-1}) \cdot P(x_t|x_{t-1})$$

- We update for evidence:

$$P(x_t|e_{1:t}) \propto_X P(x_t|e_{1:t-1}) \cdot P(e_t|x_t)$$

- The forward algorithm does both at once (and doesn’t normalize)