COL333/671: Introduction to AI
Semester I, 2022-23

Probabilistic Reasoning over Time

Rohan Paul

Outline

- Last Class
  - Probabilistic Reasoning
- This Class
  - Probabilistic Reasoning over Time
- Reference Material
  - AIMA Ch. 15

Acknowledgement

These slides are intended for teaching purposes only. Some material has been used/adapted from web sources and from slides by Doina Precup, Dorsa Sadigh, Percy Liang, Mausam, Dan Klein, Anca Dragan, Nicholas Roy and others.

Reasoning: Sequence of Observations

- Reasoning over time or space

- Several Applications
  - Monitoring a disease
  - Robot localization
  - Target Tracking
  - Speech recognition
  - User attention
  - Gesture recognition

Wildlife monitoring

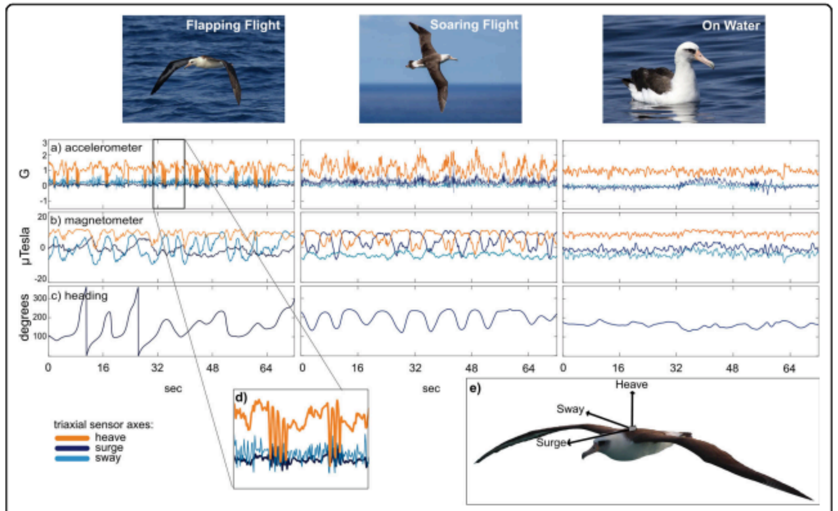


Fig. 2 Sensor data representing flapping flight, soaring flight, and 'on-water' behavior from a Laysan albatross. Distinct patterns for each behavior are observed in **a** triaxial accelerometer data, **b** triaxial magnetometer data, and **c** heading. **d** A closer look at two isolated bouts of flapping flight. **e** Magnetometer and accelerometer devices were taped centrally onto albatross backs and measured surge acceleration in the anterior-posterior axis, sway acceleration in the medio-lateral axis, and heave acceleration in the dorso-ventral axis

https://movementecologyjournal.biomedcentral.com/articles/10.1186/s40462-021-00243-z

Predicting student attrition

2.2.3 Forum Interaction Features

The forum is the primary means of student support and interaction during the course. The forum's basic software mechanisms allow us to observe the following useful features:

1. Number of threads viewed this week, where a thread can only be viewed once a day. Since most active students undertake this passive interaction it is an important metric of engagement.
2. Number of threads followed this week, which is a slightly more active sign of engagement than 1.
3. Number of upvotes given this week, indicating posts students found to be useful, which is also a more active sign of engagement.
4. Number of posts made this week. Although most students aren't active on the forum, for those who are this feature is a strong indicator of engagement and sense of community.
5. Number of replies received this week to any post previously made. This is very important as it directly correlates with how much belonging a student feels in the course.
6. Number of upvotes received this week to any post previously made. This is important for the same reasons as 5.

2.2.4 Assignment Features

Students are exposed to ungraded lecture problems that are intertwined with lecture videos, as well as graded quizzes and homeworks that assess their understanding of the material. Graded assignments are conveniently due at the end of a week. Since these types of problems carry very different weights, we define them individually as follows:

1. Cumulative percentage score on homework problems that are due at the end of this week, or have been due in previous weeks. When monitoring this value from week to week, we again get a good gauge on how far up-to-date a student is on the course.
2. Cumulative percentage score on quiz problems that are due at the end of this week, or have been due in previous weeks.
3. Cumulative percentage score on lecture problems that are available from the start of the course until this week. The difference between this and features 1 and 2 is that there is no due date for lecture problems, so a student actually has the possibility to catch up on them at any point in the course.
4. Percentage score on homework problems that are only due this week. The score that a student

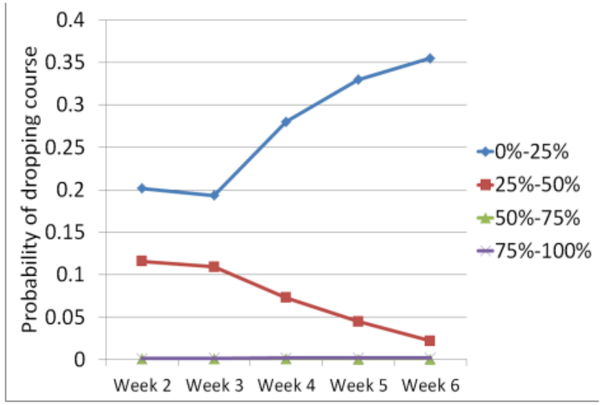
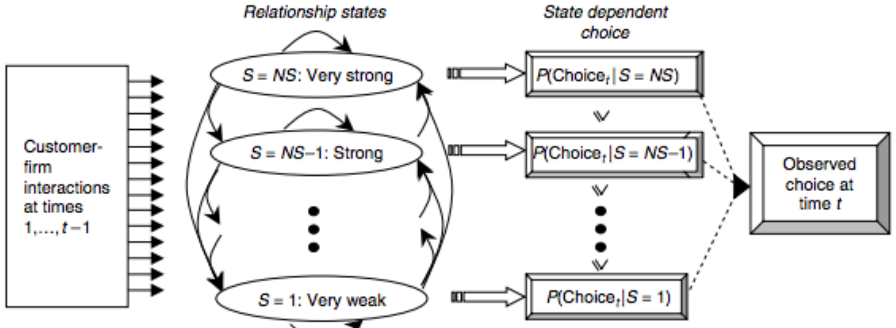


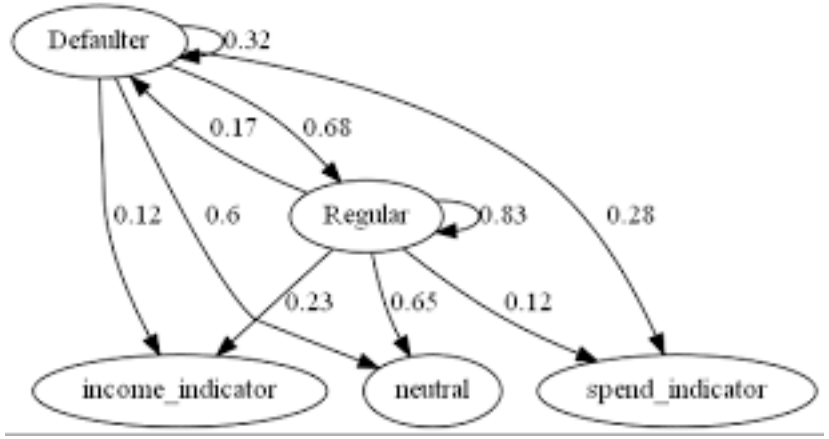
Figure 4 Attrition with time for students who view a consistent percentage of lecture videos each week. As an example, if a student is active in the course up until week 4, and views 25%-50% of lecture minutes each week during that period, their likelihood of dropping the course in week 4 is about 8%, as shown by the red line.

Predicting engagement with a university

Figure 1 A Hidden Markov Model of Customer Relationships

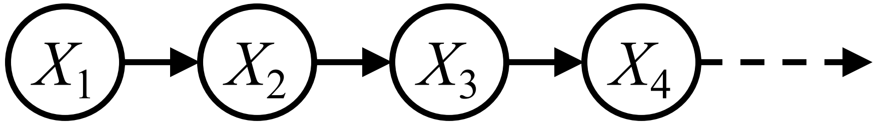


Loan monitoring



Markov Models

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



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

- **Transition probabilities** or dynamics,
  - Specify how the state evolves over time
  - Initial state probabilities
- Stationarity assumption: transition probabilities the same at all times.
- (First order) Markov Property
  - Past and future independent given the present
  - Each time step only depends on the previous

Markov Models

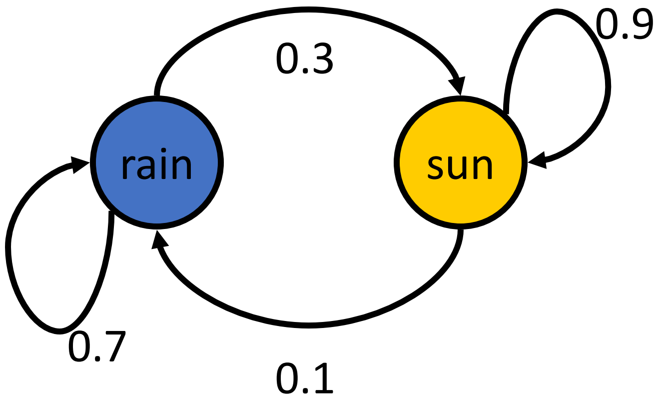
States: X = {rain, sun}

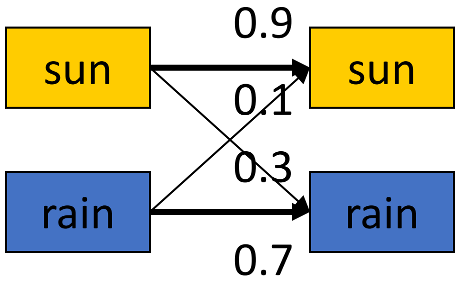
Initial distribution: 1.0 sun

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

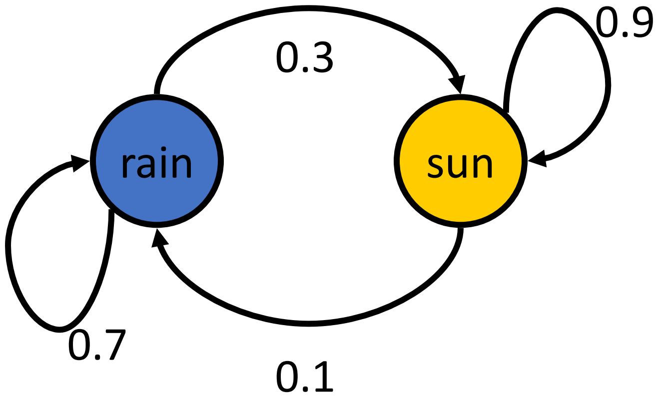
Representing the Markov model





Markov Models: Example

- Initial distribution: 1.0 sun



- What is the probability distribution after one step?

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

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

Forward Algorithm for a Markov Chain

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

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

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

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

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

Forward simulation

Forward Algorithm for a Markov Chain

- From initial observation of sun

$$\begin{pmatrix} 1.0 \\ 0.0 \end{pmatrix} \quad \begin{pmatrix} 0.9 \\ 0.1 \end{pmatrix} \quad \begin{pmatrix} 0.84 \\ 0.16 \end{pmatrix} \quad \begin{pmatrix} 0.804 \\ 0.196 \end{pmatrix} \Rightarrow \begin{pmatrix} 0.75 \\ 0.25 \end{pmatrix}$$

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

- From initial observation of rain

$$\begin{pmatrix} 0.0 \\ 1.0 \end{pmatrix} \quad \begin{pmatrix} 0.3 \\ 0.7 \end{pmatrix} \quad \begin{pmatrix} 0.48 \\ 0.52 \end{pmatrix} \quad \begin{pmatrix} 0.588 \\ 0.412 \end{pmatrix} \Rightarrow \begin{pmatrix} 0.75 \\ 0.25 \end{pmatrix}$$

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

- From yet another initial distribution $\mathrm{P}(X_1)$:

$$\begin{pmatrix} p \\ 1-p \end{pmatrix} \quad \cdots \quad \Rightarrow \begin{pmatrix} 0.75 \\ 0.25 \end{pmatrix}$$

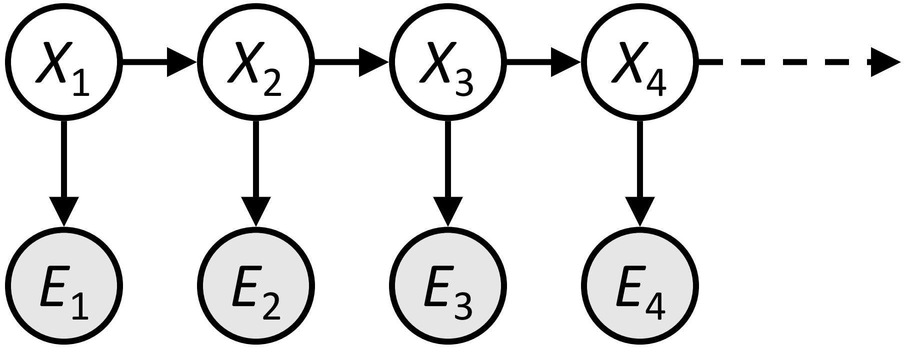
$$\mathrm{P}(X_1) \qquad\qquad \mathrm{P}(X_\infty)$$

Stationary distribution

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

Hidden Markov Models (HMMs)

- Markov Chains
  - Assume that we observe the state directly.
  - Often this is not the case. We only have noisy observations of the state.



- **Hidden** Markov Models
  - Underlying Markov chain over states X
  - You observe outputs (effects) at each time step

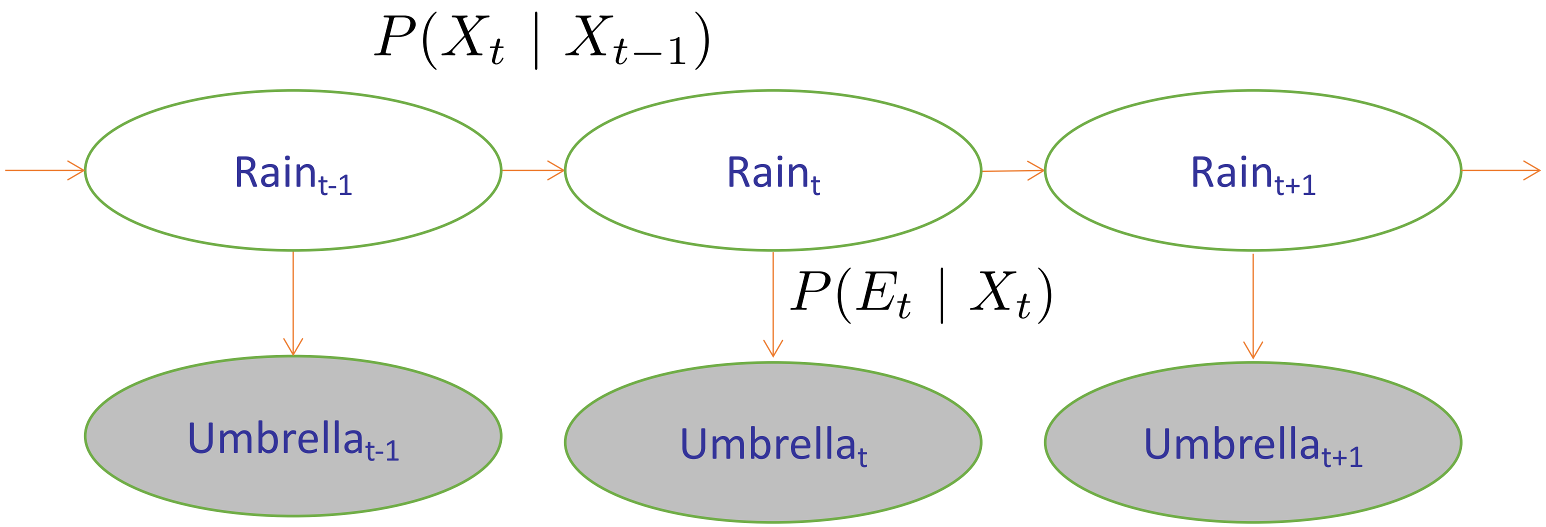
Weather HMM

The world state (rainy or sunny) is not directly observed. Instead have some observation such as a person carrying an umbrella or not.

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

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

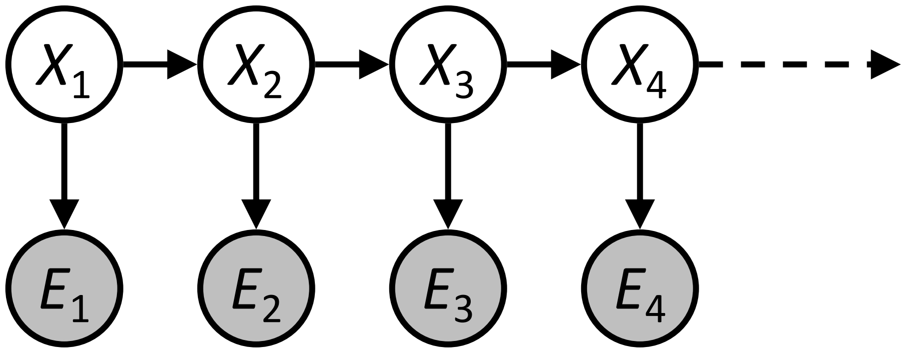
R_t	U_t	$P(U_t \mid R_t)$
+r	+u	0.9
+r	-u	0.1
-r	+u	0.2
-r	-u	0.8



HMMs – Conditional Independences

HMMs make two important independence assumptions.

- Future state depends on past states via the present state.
- The current observation is independent of all else given current state

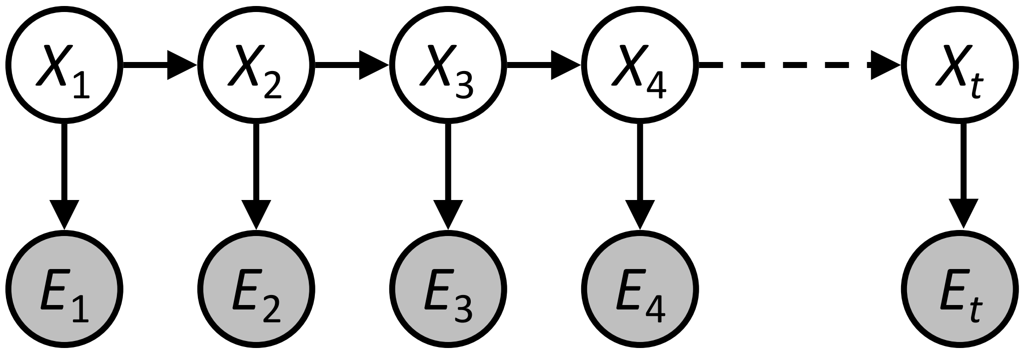


$$\mathbf{P}(\mathbf{X}_t \mid \mathbf{X}_{0:t-1}) = \mathbf{P}(\mathbf{X}_t \mid \mathbf{X}_{t-1})$$

$$\mathbf{P}(\mathbf{E}_t \mid \mathbf{X}_{0:t}, \mathbf{E}_{0:t-1}) = \mathbf{P}(\mathbf{E}_t \mid \mathbf{X}_t)$$

Filtering or 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

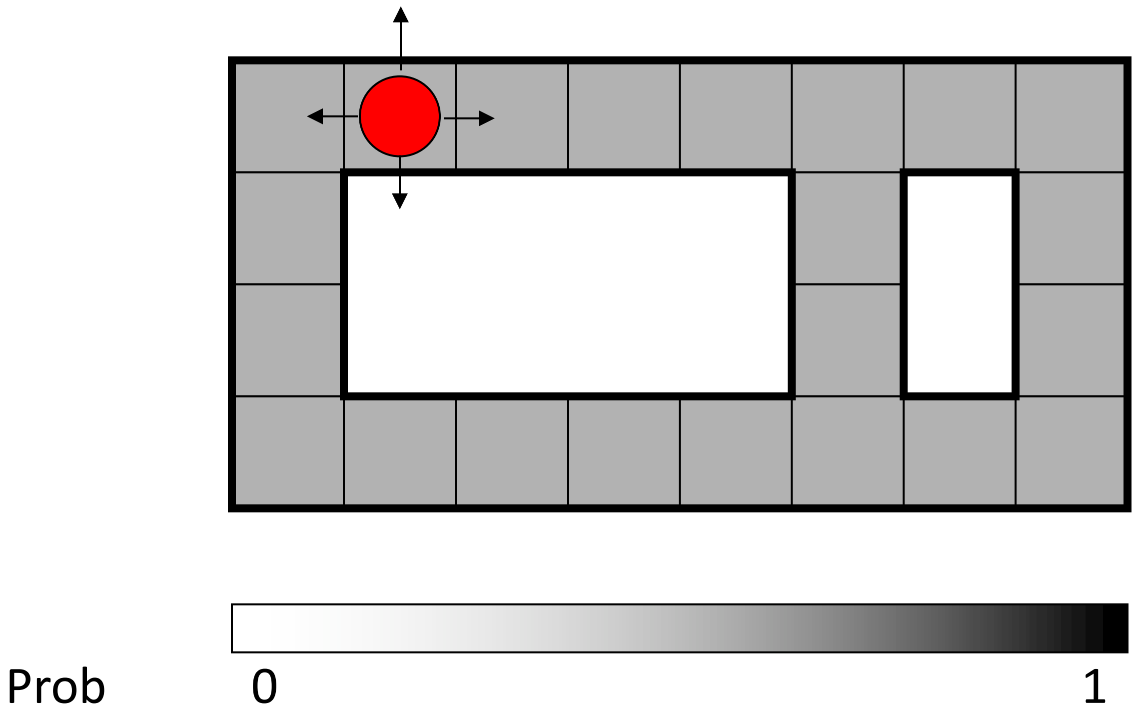


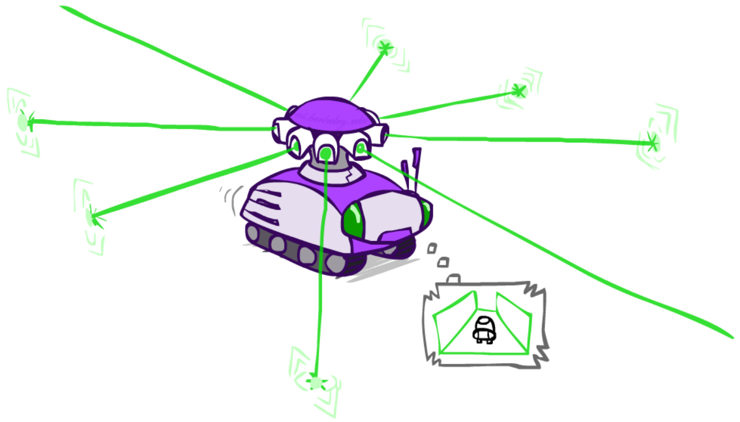
- We start with $B_1(X)$ in an initial setting, usually uniform

- As time passes, or we get observations, we update $B(X)$

Example: Robot Localization

Robot can take actions N, S, E, W
Detects walls from its sensors



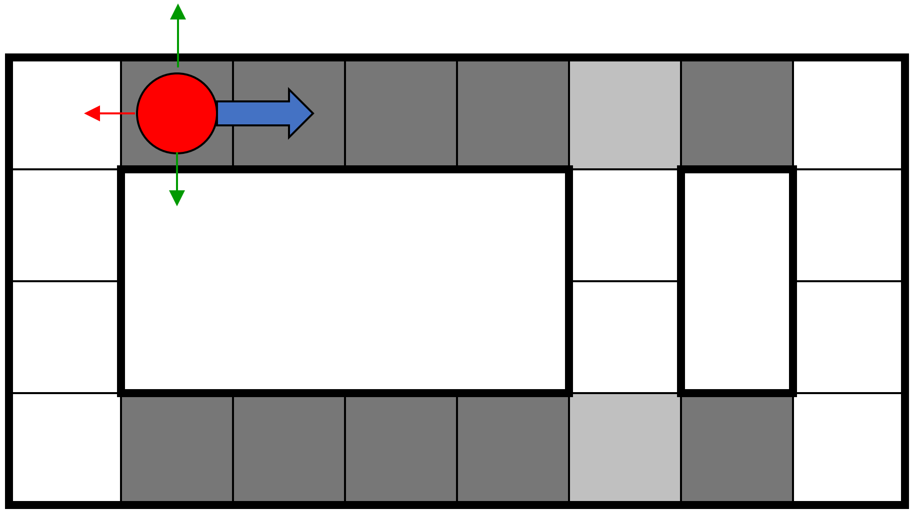


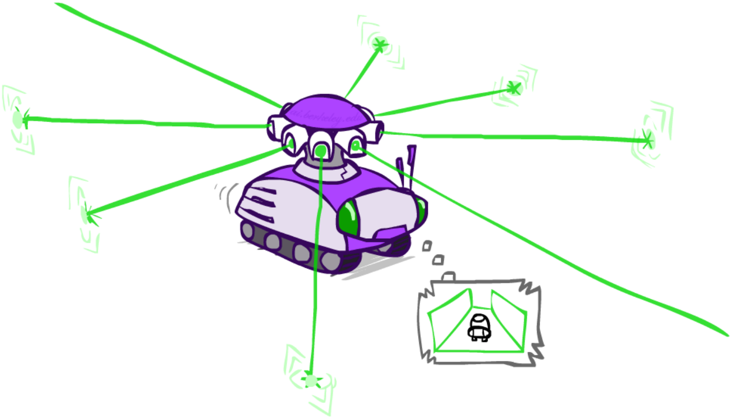
t=0

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



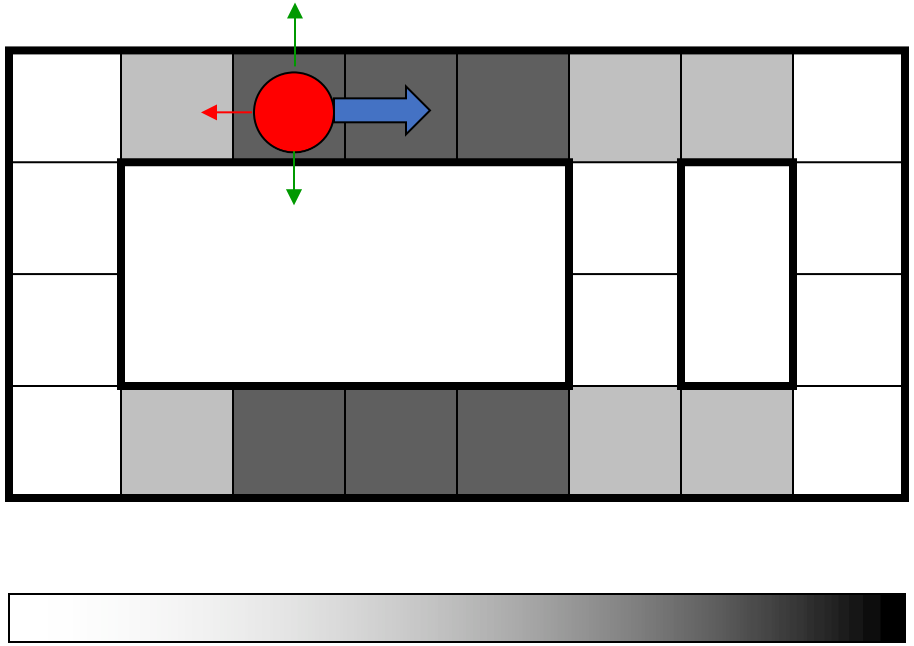


Prob 0 1

t=1

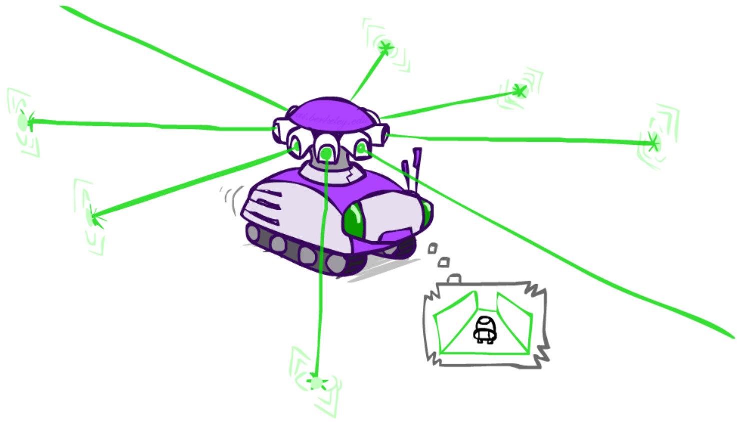
Lighter grey: was possible to get the reading, but less likely b/c required 1 mistake

Example: Robot Localization

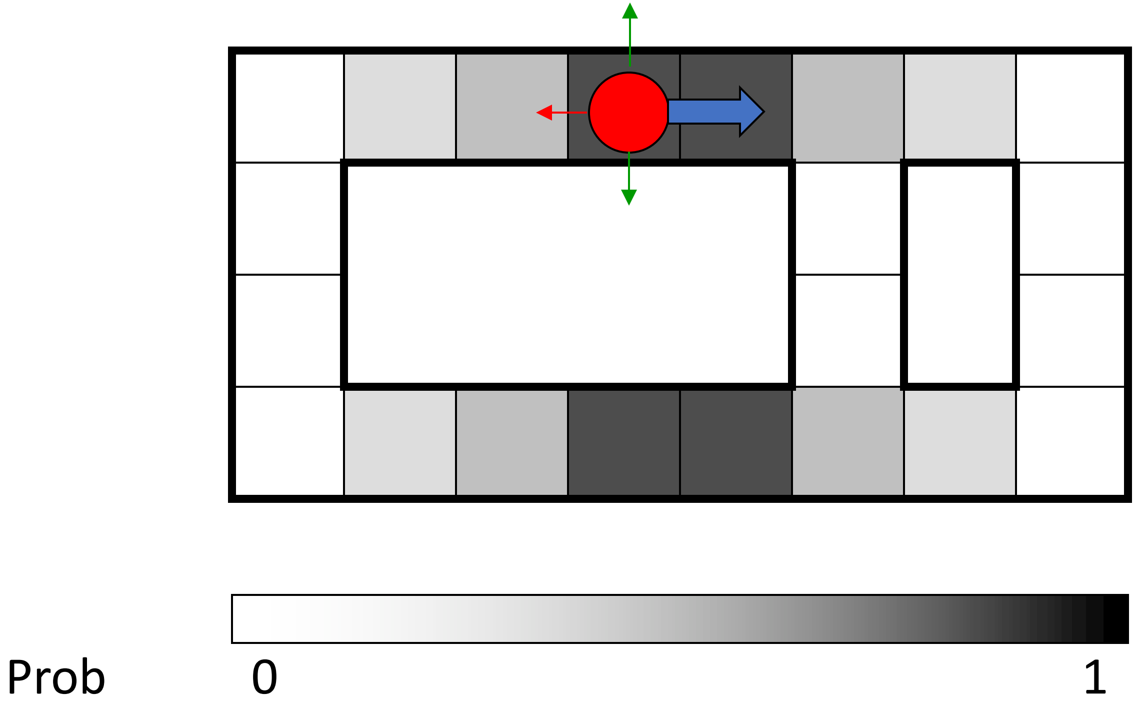


Prob 0 1

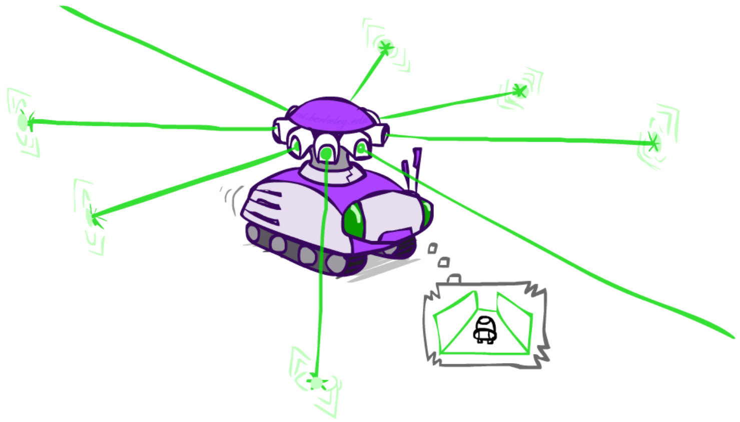
t=2



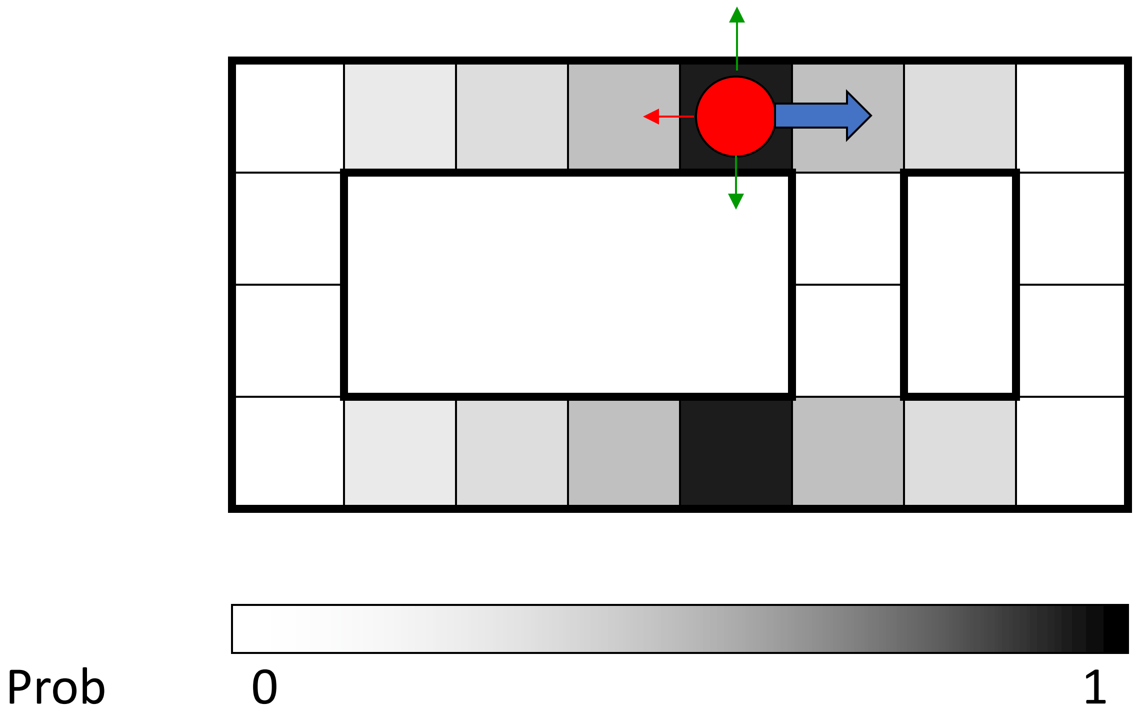
Example: Robot Localization



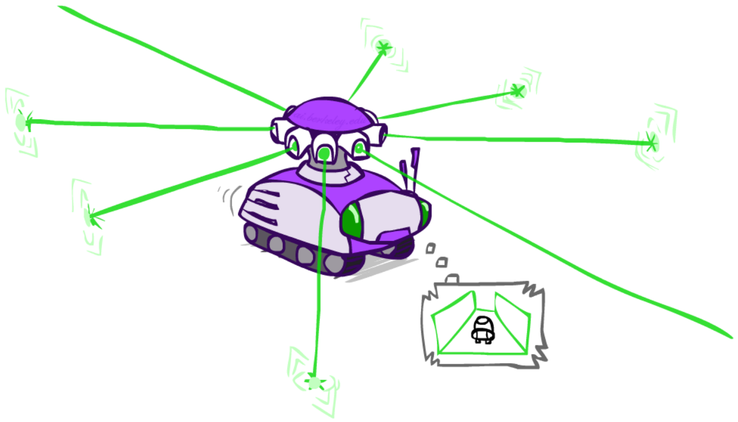
t=3



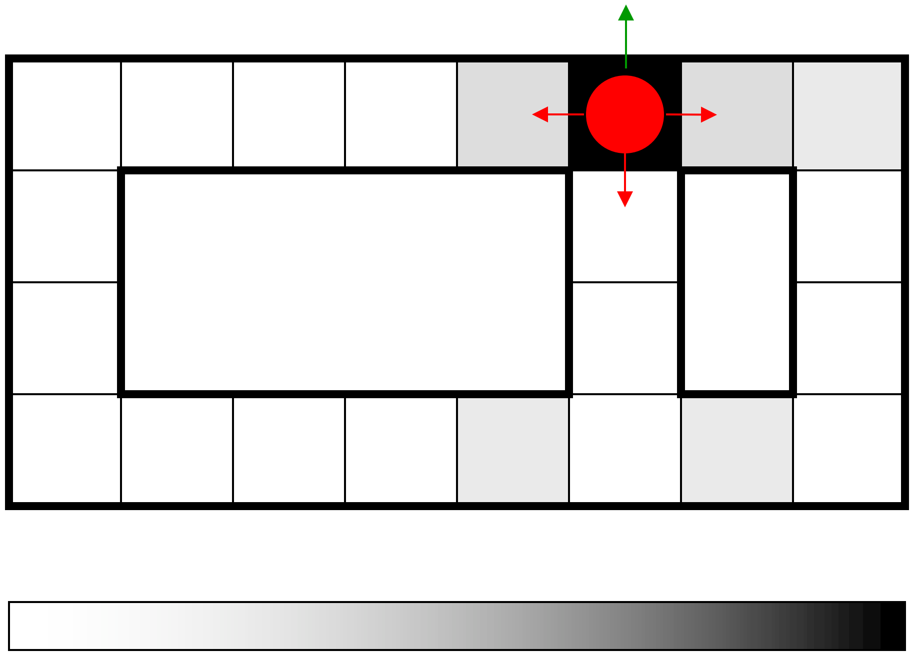
Example: Robot Localization



t=4

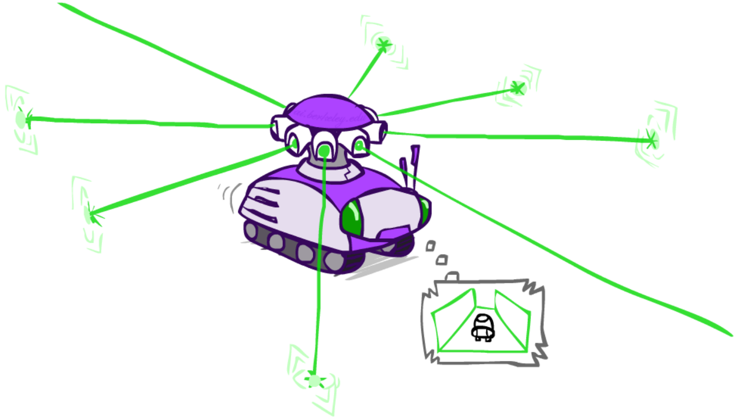


Example: Robot Localization



Prob 0 1

t=5



Inference: Estimate State Given Evidence

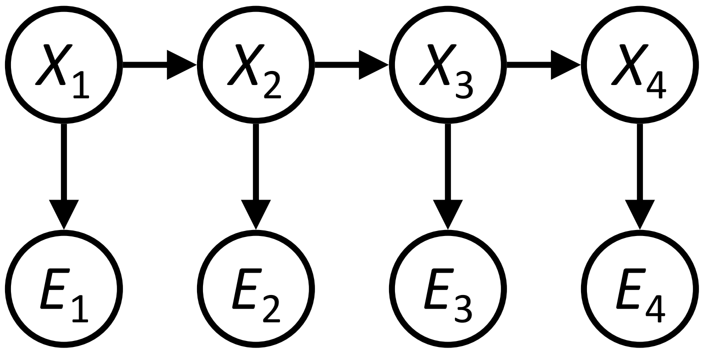
- We are given evidence at each time and want to know

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

- Approach: start with $P(X_1)$ and derive B_t in terms of B_{t-1}
  - Equivalently, derive B_{t+1} in terms of B_t

- Two Steps:
  - Passage of time
  - Evidence incorporation

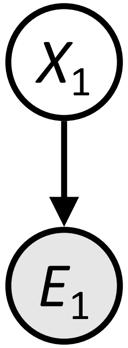
$B(X_t) = P(X_t|e_{1:t})$ $B'(X_{t+1})$



$B(X_{t+1})$

Estimating State Given Evidence: Base Cases

- Evidence incorporation
  - Incorporating noisy observations of the state.

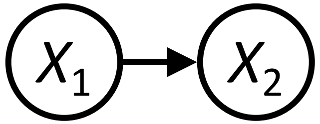


$$P(X_1|e_1)$$

$$P(X_1|e_1) = \frac{P(X_1, e_1)}{\sum_{x_1} P(x_1, e_1)}$$

$$P(X_1|e_1) = \frac{P(e_1|X_1)P(X_1)}{\sum_{x_1} P(e_1|x_1)P(x_1)}$$

- Passage of time
  - The system state at the next time step given transition model



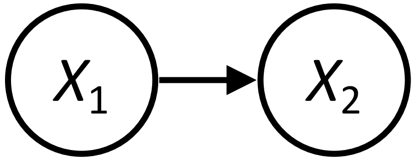
$$P(X_2)$$

$$P(X_2) = \sum_{x_1} P(x_1, X_2)$$

$$P(X_2) = \sum_{x_1} P(X_2|x_1)P(x_1)$$

Next, perform these two computations repeatedly over each time step

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:

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

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

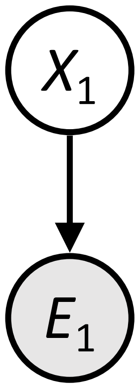
$$= \sum_{x_t} P(X_{t+1}|x_t)P(x_t|e_{1:t})$$

Basic idea: the beliefs get "pushed" through the transitions

Incorporating Observations

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:

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

$$\propto_{X_{t+1}} P(X_{t+1}, e_{t+1}|e_{1:t})$$

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

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

View it as a "correction" of the belief using the observation

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

Inference: Weather HMM

B(+r) = 0.5
B(-r) = 0.5

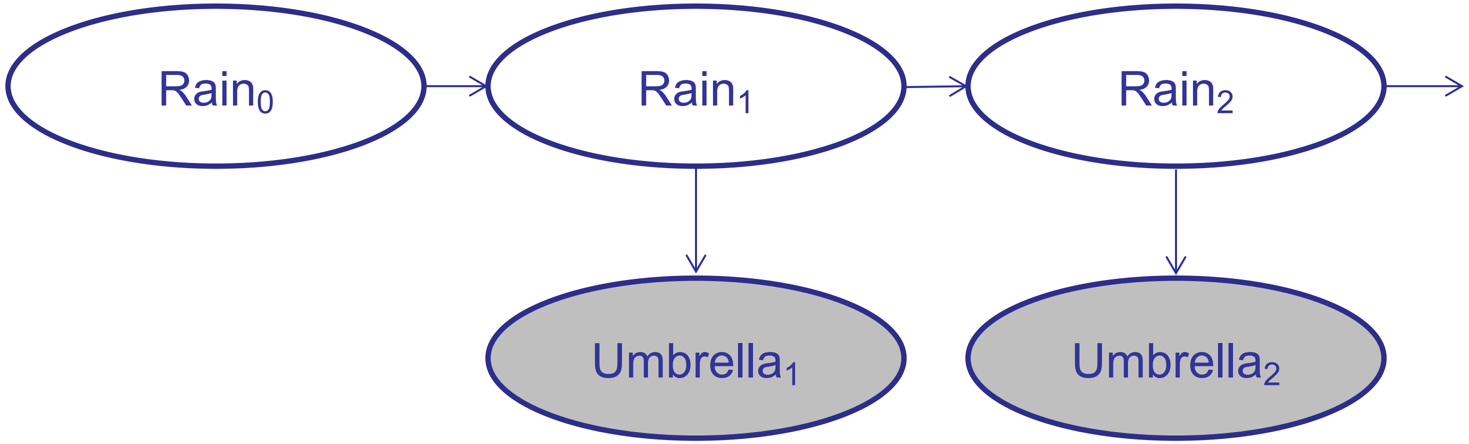
B'(+r) = 0.5
B'(-r) = 0.5

B(+r) = 0.818
B(-r) = 0.182

B'(+r) = 0.627
B'(-r) = 0.373

B(+r) = 0.883
B(-r) = 0.117

Passage of time and correction at each stage.



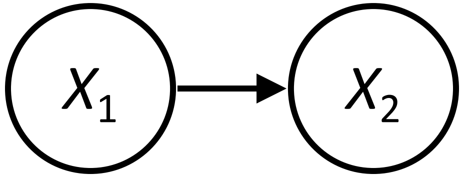
R_t	R_{t+1}	$P(R_{t+1}\|R_t)$
+r	+r	0.7
+r	-r	0.3
-r	+r	0.3
-r	-r	0.7

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

Online Belief Updates: Inference over Time

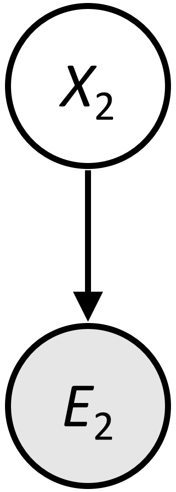
- Every time step, we start with current P(X | evidence)
- We update for time:

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



- We update for evidence:

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



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

$$
\begin{aligned}
P(x_t | e_{1:t}) &\propto_{X_t} P(x_t, e_{1:t}) \\
&= \sum_{x_{t-1}} P(x_{t-1}, x_t, e_{1:t}) \\
&= \sum_{x_{t-1}} P(x_{t-1}, e_{1:t-1}) P(x_t | x_{t-1}) P(e_t | x_t) \\
&= P(e_t | x_t) \sum_{x_{t-1}} P(x_t | x_{t-1}) P(x_{t-1}, e_{1:t-1})
\end{aligned}
$$

Normalization can be at each step if the exact likelihood is needed at each step or at the end.

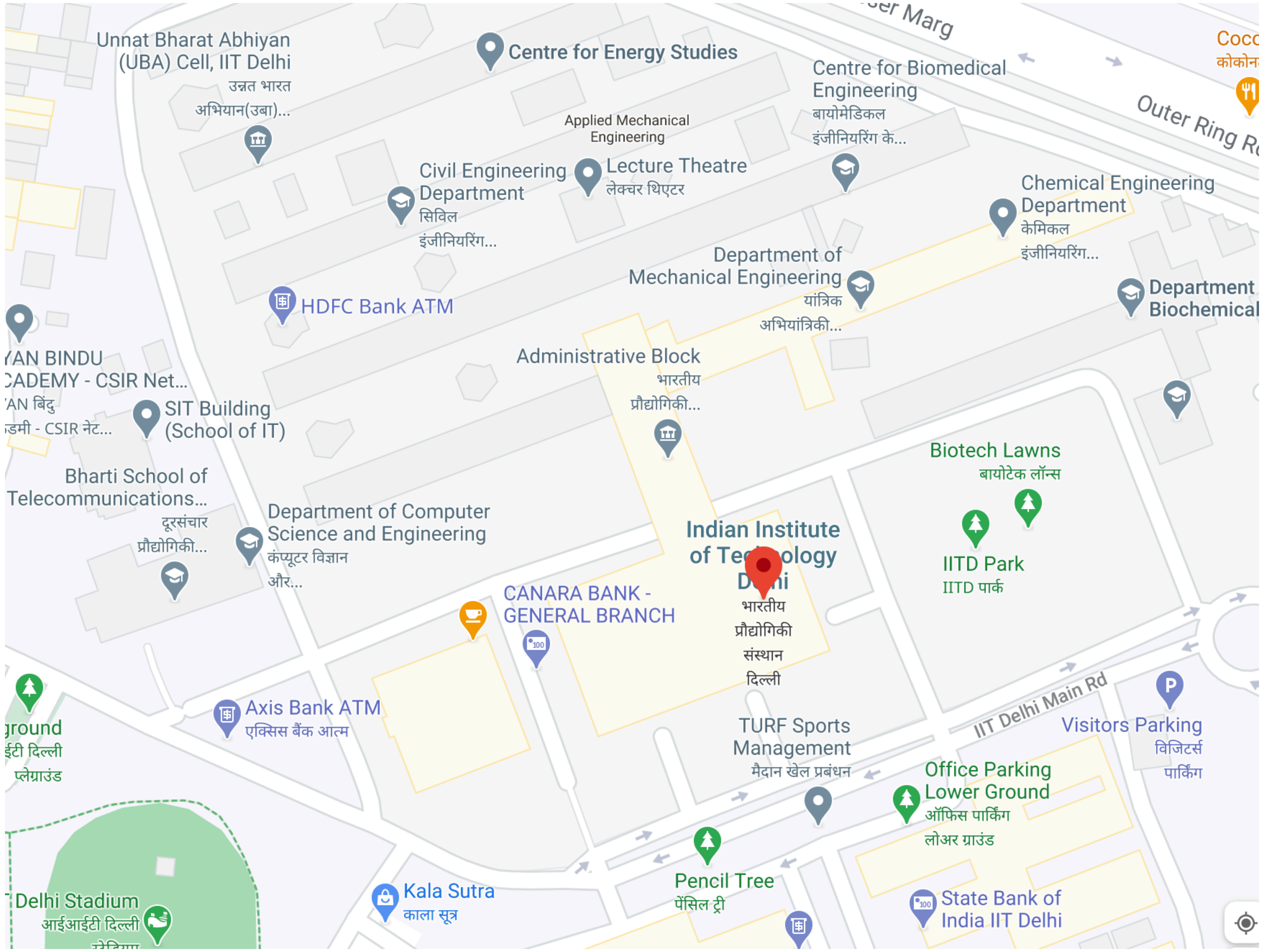
Large Number of States

A grid over a large space can lead to a large state space.

Imagine tracking a car at a city scale.

The grid size will be very large!

Difficult to run the forward algorithm.



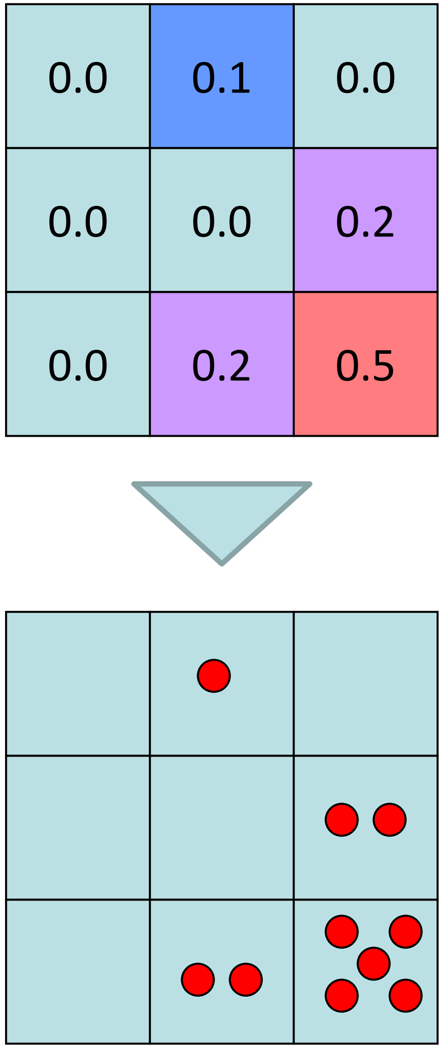
Particle Filtering

Problem: Sometimes |X| is too big to use exact inference

- |X| may be too big to even store B(X)
- E.g. X is continuous (though here we focus on the discrete case)

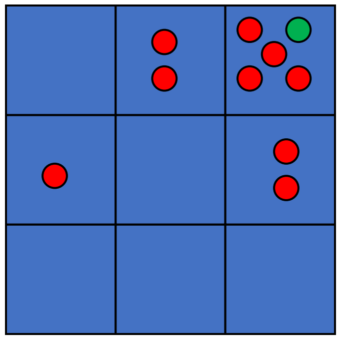
Solution: approximate inference

- Track samples of X, not all values.
- Samples are called “particles”
- Time spent per step is linear in the number of samples
- Keep the list of particles in memory, not states
- Larger the number of particles, the better is the approximation.



Representation: Particles

- Our representation of P(X) is now a list of N particles (samples)
  - Generally, N << |X|
- P(x) approximated by number of particles with value x
  - Several x can have P(x) = 0. Note that (3,3) has half the number of particles.
  - Larger the number of particles, better is the approximation.



Particles:
(3,3)
(2,3)
(3,3)
(3,2)
(3,3)
(3,2)
(1,2)
(3,3)
(3,3)
(2,3)

Representation: Passage of Time

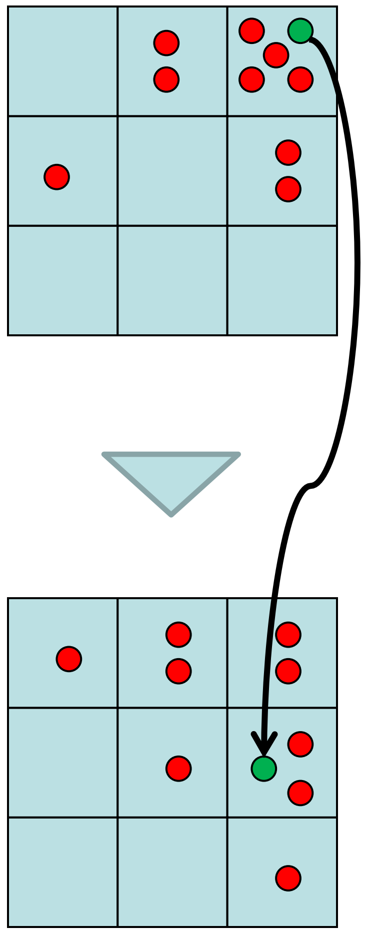
Each particle is moved by sampling its next position from the transition model

$$x' = \text{sample}(P(X'|x))$$

- Perform simulation or sampling
    - The samples' frequencies reflect the transition probabilities
- In the example, most samples move clockwise, but some move in another direction or stay in place.
    - This is an outcome of the probabilistic transition model.

Particles:
(3,3)
(2,3)
(3,3)
(3,2)
(3,3)
(3,2)
(1,2)
(3,3)
(3,3)
(2,3)

Particles:
(3,2)
(2,3)
(3,2)
(3,1)
(3,3)
(3,2)
(1,3)
(2,3)
(3,2)
(2,2)



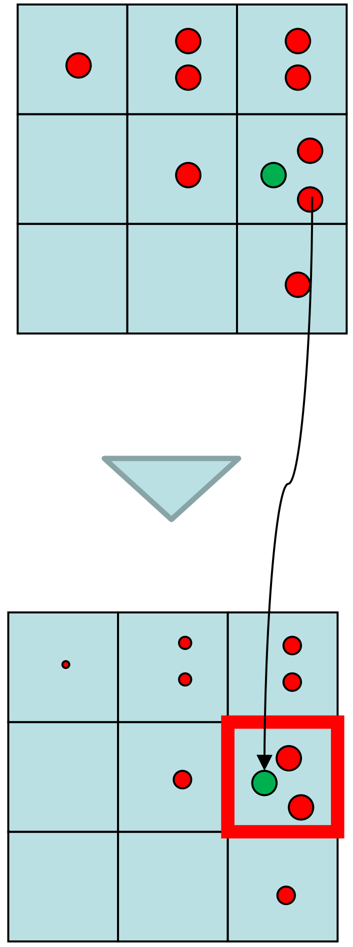
Representation: Incorporate Evidence

- As seen previously, incorporating evidence adjusts or weighs the probabilities.
- Attach a weight to each sample.
- Weigh the samples based on the likelihood of the evidence.

$$w(x) = P(e|x)$$

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

Particles:
(3,2)
(2,3)
(3,2)
(3,1)
(3,3)
(3,2)
(1,3)
(2,3)
(3,2)
(2,2)



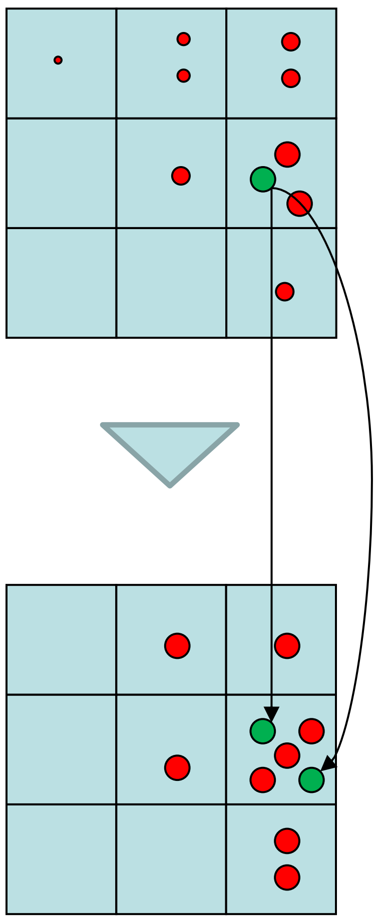
Particles:
(3,2) w=.9
(2,3) w=.2
(3,2) w=.9
(3,1) w=.4
(3,3) w=.4
(3,2) w=.9
(1,3) w=.1
(2,3) w=.2
(3,2) w=.9
(2,2) w=.4

Representation: Resample

- Rather than tracking weighted samples, we resample
- N times, we choose from our weighted sample distribution (i.e. draw with replacement)
- Now the update is complete for this time step, continue with the next one

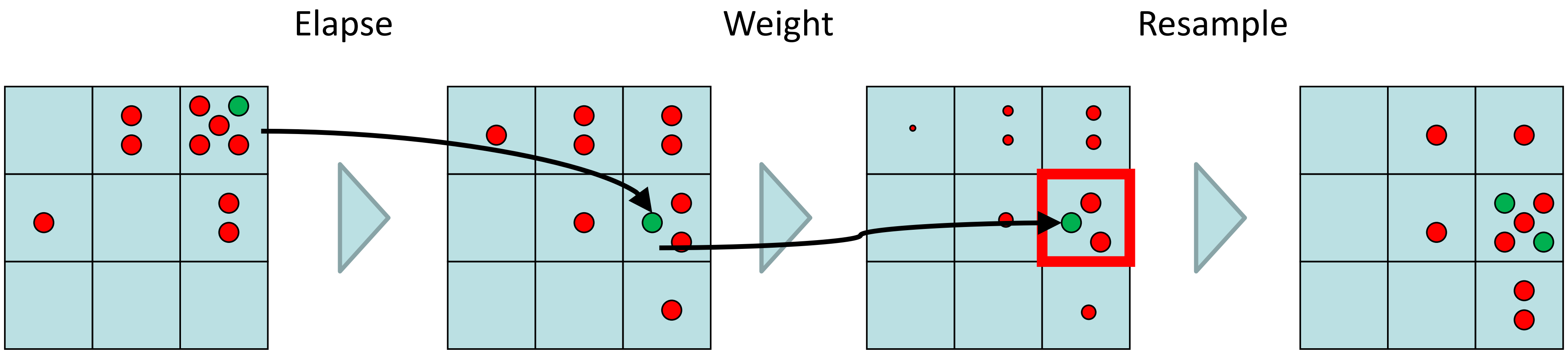
Particles:
(3,2) w=.9
(2,3) w=.2
(3,2) w=.9
(3,1) w=.4
(3,3) w=.4
(3,2) w=.9
(1,3) w=.1
(2,3) w=.2
(3,2) w=.9
(2,2) w=.4

(New) Particles:
(3,2)
(2,2)
(3,2)
(2,3)
(3,3)
(3,2)
(1,3)
(2,3)
(3,2)
(3,2)

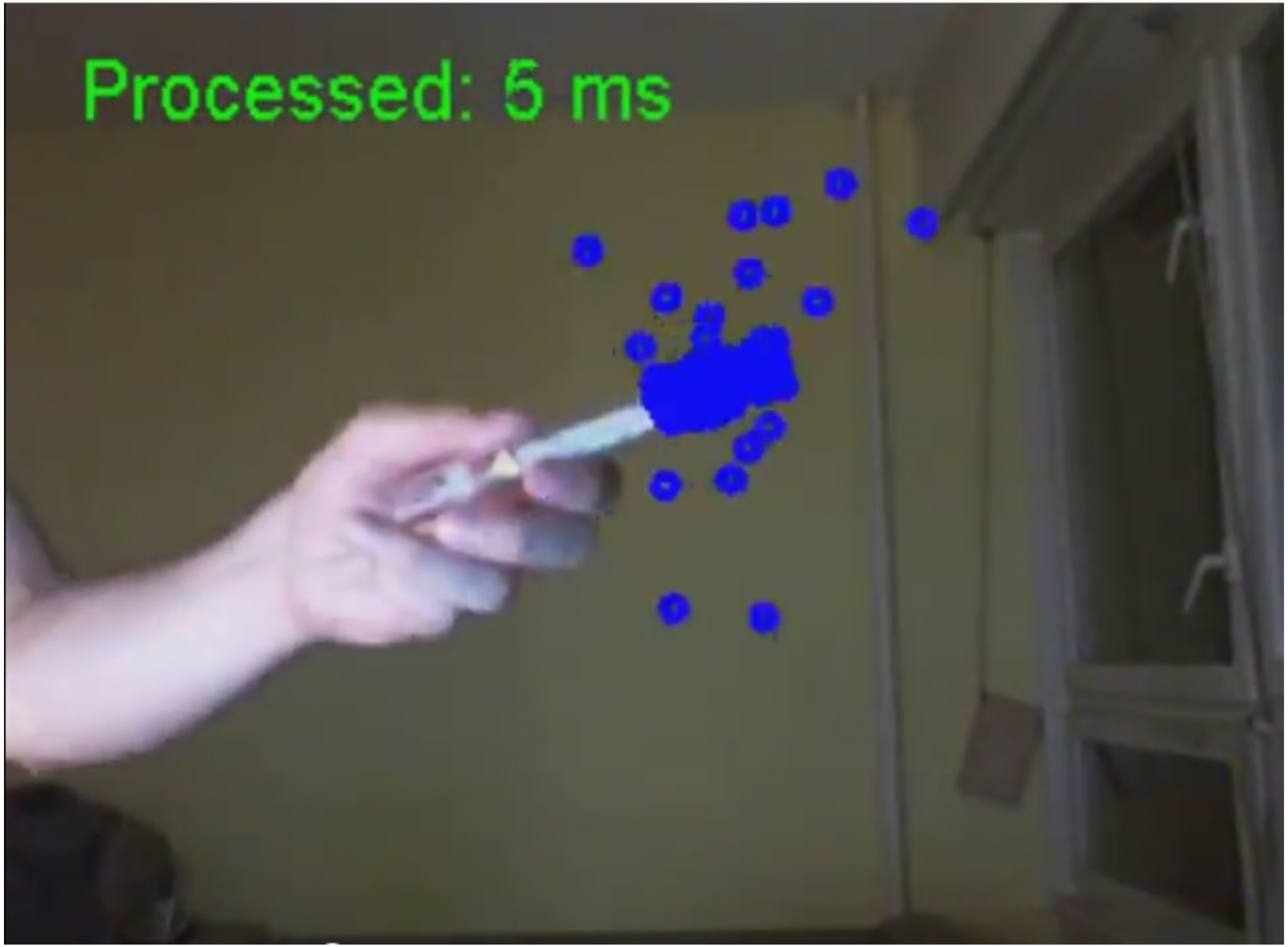


Representation: Particles

Particles: track samples of states rather than an explicit distribution



Particles:	Particles:	Particles:	(New) Particles:
(3,3)	(3,2)	(3,2) w=.9	(3,2)
(2,3)	(2,3)	(2,3) w=.2	(2,2)
(3,3)	(3,2)	(3,2) w=.9	(3,2)
(3,2)	(3,1)	(3,1) w=.4	(2,3)
(3,3)	(3,3)	(3,3) w=.4	(3,3)
(3,2)	(3,2)	(3,2) w=.9	(3,2)
(1,2)	(1,3)	(1,3) w=.1	(1,3)
(3,3)	(2,3)	(2,3) w=.2	(2,3)
(3,3)	(3,2)	(3,2) w=.9	(3,2)
(2,3)	(2,2)	(2,2) w=.4	(3,2)

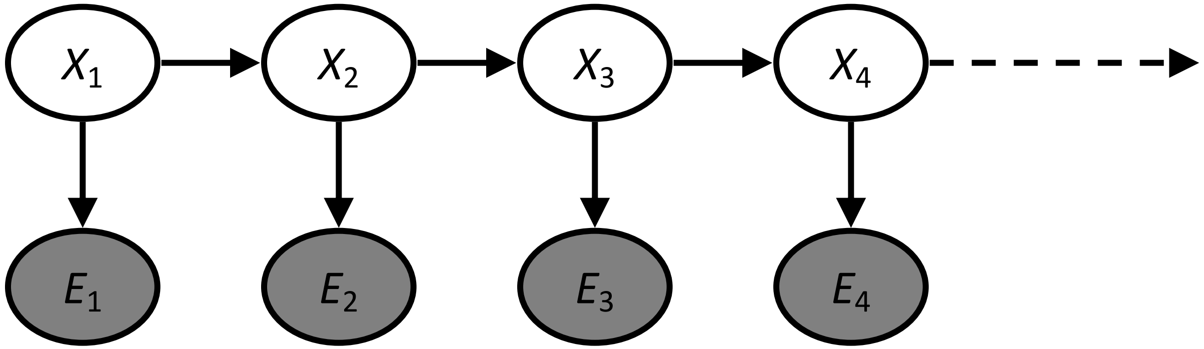


Application: tracking of a red pen. The blue dots indicate the estimated positions.
Video: https://www.youtube.com/watch?v=SV6CmEha51k

Most Likely Explanation Queries

HMMs defined by

- States X
- Observations E
- Initial distribution: $P(X_1)$
- Transitions: $P(X|X_{-1})$
- Emissions: $P(E|X)$



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

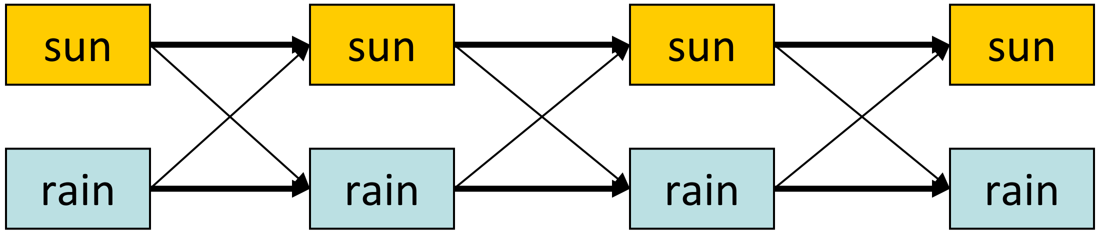
Problem: Most-likely Explanation

Determine the most likely sequence of states given all the evidence.

Solution: the Viterbi algorithm

State Trellis

State trellis: graph of states and transitions over time



X_1 X_2 $\cdots$ X_N

Each arc represents some transition $x_{t-1} \rightarrow x_t$

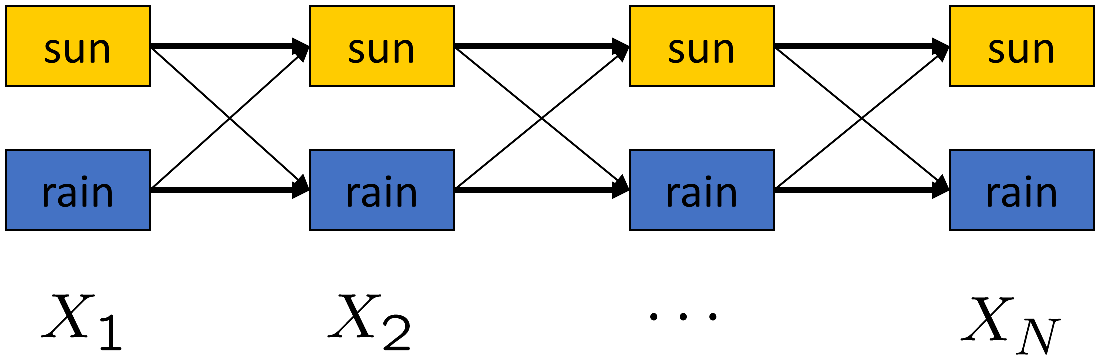
Each arc has weight $P(x_t|x_{t-1})P(e_t|x_t)$

Each path is a sequence of states

The product of weights on a path is that sequence's probability along with the evidence

The most likely explanation query – is like finding the best path in this structure.

Viterbi Algorithm



Forward Algorithm (Sum)

$$f_t[x_t] = P(x_t, e_{1:t})$$

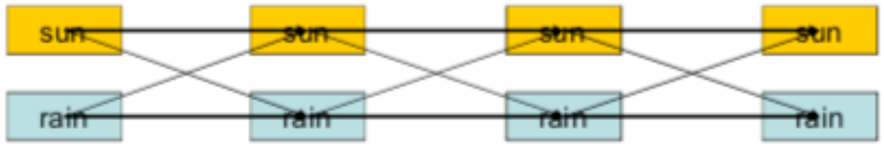
$$= P(e_t|x_t) \sum_{x_{t-1}} P(x_t|x_{t-1}) f_{t-1}[x_{t-1}]$$

Viterbi Algorithm (Max)

$$m_t[x_t] = \max_{x_{1:t-1}} P(x_{1:t-1}, x_t, e_{1:t})$$

$$= P(e_t|x_t) \max_{x_{t-1}} P(x_t|x_{t-1}) m_{t-1}[x_{t-1}]$$

Viterbi Algorithm



$$x^*_{1:T} = \arg\max_{x_{1:T}} P(x_{1:T}|e_{1:T}) = \arg\max_{x_{1:T}} P(x_{1:T}, e_{1:T})$$

$$\begin{aligned}
m_t[x_t] &= \max_{x_{1:t-1}} P(x_{1:t-1}, x_t, e_{1:t}) \\
&= \max_{x_{1:t-1}} P(x_{1:t-1}, e_{1:t-1}) P(x_t|x_{t-1}) P(e_t|x_t) \\
&= P(e_t|x_t) \max_{x_{t-1}} P(x_t|x_{t-1}) \max_{x_{1:t-2}} P(x_{1:t-1}, e_{1:t-1}) \\
&= P(e_t|x_t) \max_{x_{t-1}} P(x_t|x_{t-1}) m_{t-1}[x_{t-1}]
\end{aligned}$$