Markov Models

Uncertainty and Time

- Often, we want to reason about a ***sequence*** of observations where the state of the underlying system is ***changing***
  - Speech recognition
  - Robot localization
  - User attention
  - Medical monitoring
  - Global climate
- Need to introduce time into our models

Markov Models (aka Markov chain/process)

$$\boxed{X \to Y \to Z}$$

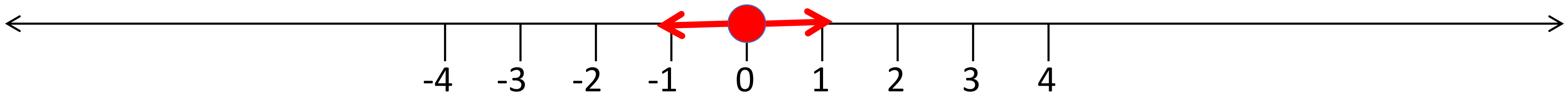
$$P(X_t = x \mid X_{t-1} = y) = f(x, y)$$

$$X_0 \to X_1 \to X_2 \to X_3 \dashrightarrow$$

$P(X_0)$ $\qquad$ $P(X_t \mid X_{t-1})$

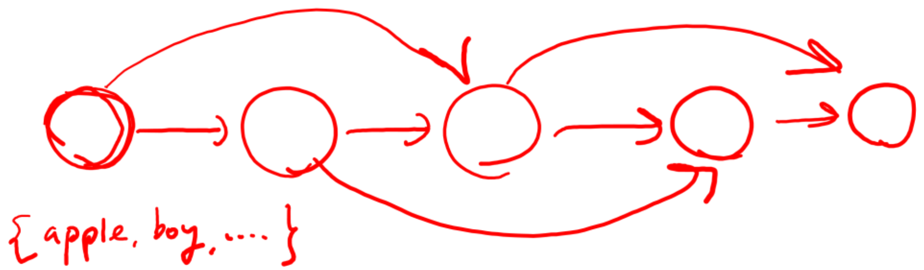
- Value of X at a given time is called the **state**
- The **transition model** $P(X_t \mid X_{t-1})$ specifies how the state evolves over time
- **Stationarity** assumption: transition probabilities are the same at all times
- **Markov** assumption: "future is independent of the past given the present"
  - X_{t+1} is independent of $X_0, \ldots, X_{t-1}$ given X_t

Example: Random walk in one dimension



- State: location on the unbounded integer line
- Initial probability: starts at 0
- Transition model: $P(X_t = k | X_{t-1} = k \pm 1) = 0.5$
- Applications: particle motion in crystals, stock prices, etc.

Example: n-gram models

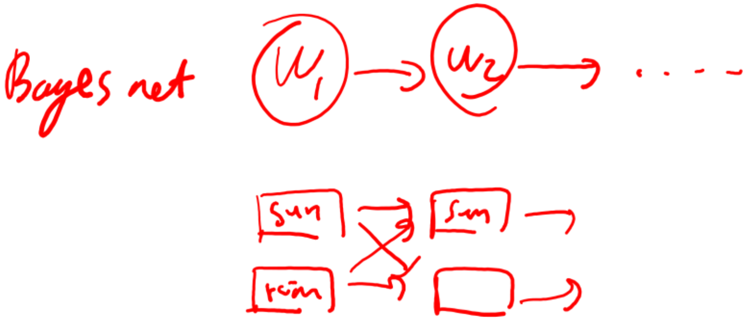


- State: word at position t in text (can also build letter n-grams)
- Transition model (probabilities come from empirical frequencies):
  - Unigram (zero-order): $P(Word_t = i)$
    - "logical are as are confusion a may right tries agent goal the was . . ."
  - Bigram (first-order): $P(Word_t = i \mid Word_{t-1} = j)$
    - "systems are very similar computational approach would be represented . . ."
  - Trigram (second-order): $P(Word_t = i \mid Word_{t-1} = j, Word_{t-2} = k)$
    - "planning and scheduling are integrated the success of naive bayes model is . . ."
- Applications: text classification, spam detection, author identification, language classification, speech recognition

Example: Web browsing

- State: URL visited at step t
- Transition model:
  - With probability p, choose an outgoing link at random
  - With probability $(1-p)$, choose an arbitrary new page
- Question: What is the ***stationary distribution*** over pages?
  - I.e., if the process runs forever, what fraction of time does it spend in any given page?
- Application: Google page rank

Example: Weather



- States {rain, sun}

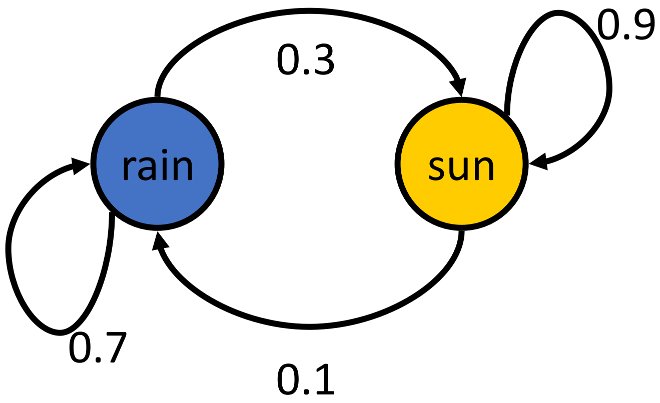
- Initial distribution $P(X_0)$

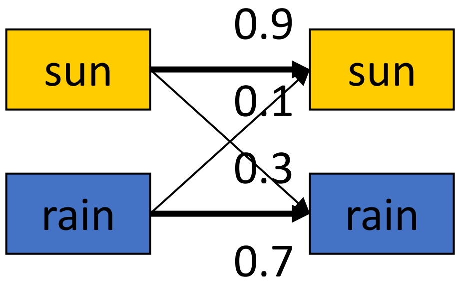
$P(X_0)$	
sun	rain
0.5	0.5

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

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

Two ways to represent Markov chains





Weather prediction

$P(X_{t-1}) \xrightarrow{\text{join}} P(X_{t-1}, X_t) \xrightarrow{\text{margin}} P(X_t)$

- Time 0: <0.5,0.5>

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

- What is the weather like at time 1?

$$P(X_1) = \sum_{x_0} P(X_1, X_0 = x_0)$$
$$= \sum_{x_0} P(X_0 = x_0)\, P(X_1 \mid X_0 = x_0)$$
$$= 0.5\langle 0.9, 0.1\rangle + 0.5\langle 0.3, 0.7\rangle = \langle 0.6, 0.4\rangle$$

Weather prediction, contd.

- Time 1: <0.6,0.4>

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

- What is the weather like at time 2?

$$P(X_2) = \sum_{x_1} P(X_2, X_1{=}x_1)$$
$$= \sum_{x_1} P(X_1{=}x_1)\, P(X_2 \mid X_1{=}x_1)$$
$$= 0.6\langle 0.9,0.1\rangle + 0.4\langle 0.3,0.7\rangle = \langle 0.66,0.34\rangle$$

Weather prediction, contd.

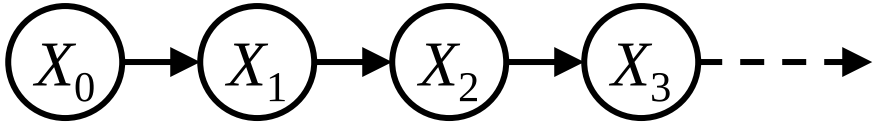
- Time 2: <0.66,0.34>

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

- What is the weather like at time 3?

$$P(X_3) = \sum_{x_2} P(X_3, X_2 = x_2)$$
$$= \sum_{x_2} P(X_2 = x_2)\, P(X_3 \mid X_2 = x_2)$$
$$= 0.66\langle 0.9, 0.1\rangle + 0.34\langle 0.3, 0.7\rangle = \langle 0.696, 0.304\rangle$$

Forward algorithm (simple form)



$P(X_0)$ $\quad$ $P(X_t \mid X_{t-1})$

What is the state at time t?

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

Forward algorithm in Matrices

- What is the weather like at time 2?
  - $P(X_2) = 0.6\langle 0.9,0.1\rangle + 0.4\langle 0.3,0.7\rangle = \langle 0.66,0.34\rangle$
- In matrix-vector form:
  - $P(X_2) = \begin{pmatrix} 0.9 & 0.3 \\ 0.1 & 0.7 \end{pmatrix} \underbrace{\begin{pmatrix} 0.6 \\ 0.4 \end{pmatrix}}_{P(X_1)} = \begin{pmatrix} 0.66 \\ 0.34 \end{pmatrix}$

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

Stationary Distributions

- The limiting distribution is called the ***stationary distribution*** P_∞ of the chain
- It satisfies $P_\infty = P_{\infty+1} = T^{\mathrm{T}} P_\infty$

  Stationary distribution is <0.75,0.25> ***regardless of starting distribution***

sum

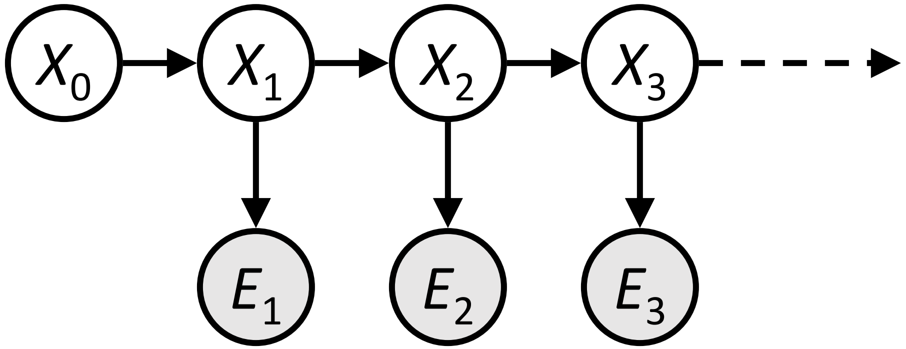
$$\begin{pmatrix} 0.9 & 0.3 \\ 0.1 & 0.7 \end{pmatrix} \begin{pmatrix} p \\ \underline{1-p} \end{pmatrix} = \begin{pmatrix} p \\ 1-p \end{pmatrix}$$

Hidden Markov Models

Hidden Markov Models

- Usually the true state is not observed directly

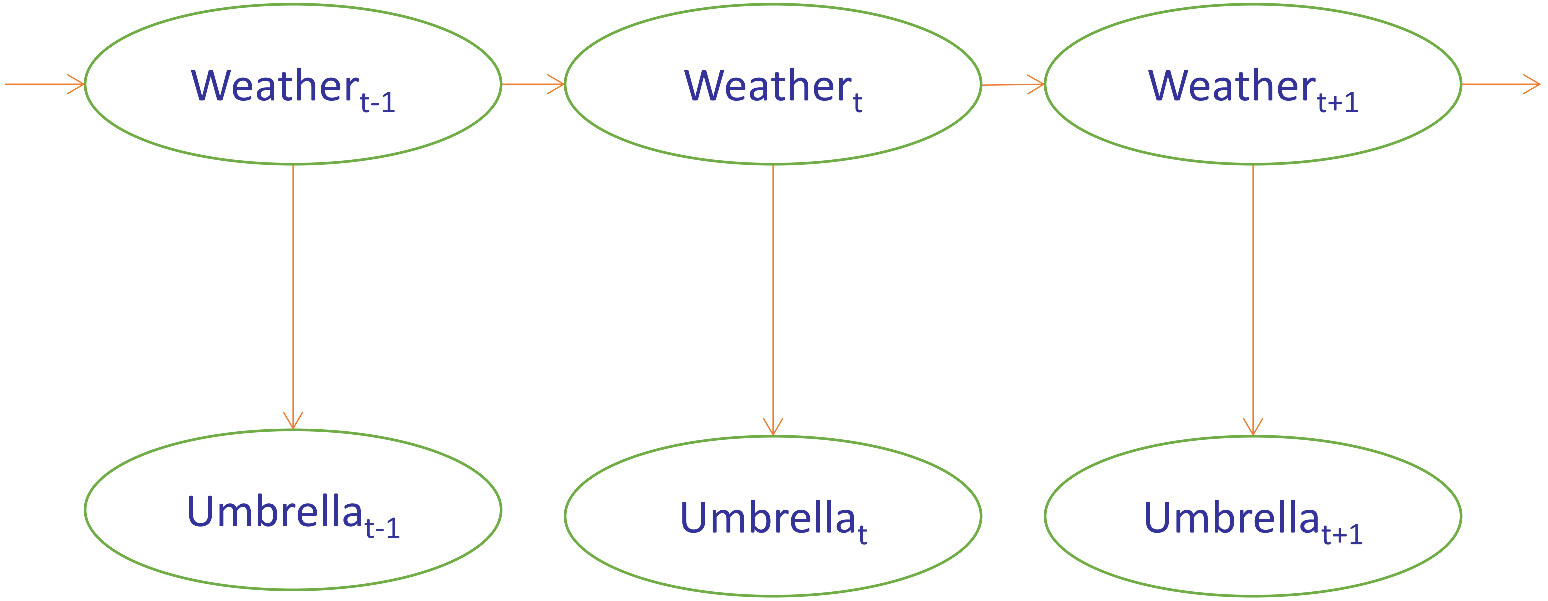
- Hidden Markov models (HMMs)
  - Underlying Markov chain over states X
  - You observe evidence E at each time step
  - X_t is a single discrete variable; E_t may be continuous and may consist of several variables



Example: Weather HMM

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

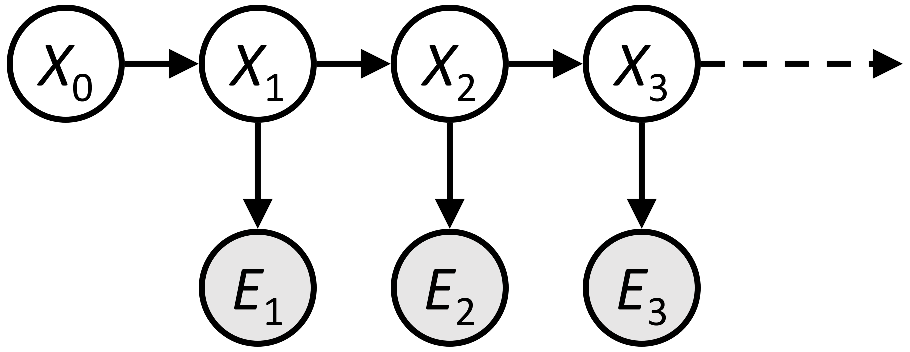
- An HMM is defined by:
  - Initial distribution: $P(X_0)$
  - Transition model: $P(X_t \mid X_{t-1})$
  - Sensor model: $P(E_t \mid X_t)$



W_t	$P(U_t \mid W_t)$	
	true	false
sun	0.2	0.8
rain	0.9	0.1

HMM as probability model

- Joint distribution for Markov model: $P(X_0,\ldots, X_T) = P(X_0) \prod_{t=1:T} P(X_t \mid X_{t-1})$
- Joint distribution for hidden Markov model:
  $P(X_0,E_0,X_1,E_1,\ldots, X_T,E_T) = P(X_0) \prod_{t=1:T} P(X_t \mid X_{t-1})\, P(E_t \mid X_t)$
- Future states are independent of the past given the present
- Current evidence is independent of everything else given the current state
- Are evidence variables independent of each other?



Real HMM Examples

apple

- 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)
- Molecular biology:
  - Observations are nucleotides ACGT
  - States are coding/non-coding/start/stop/splice-site etc.

Inference tasks

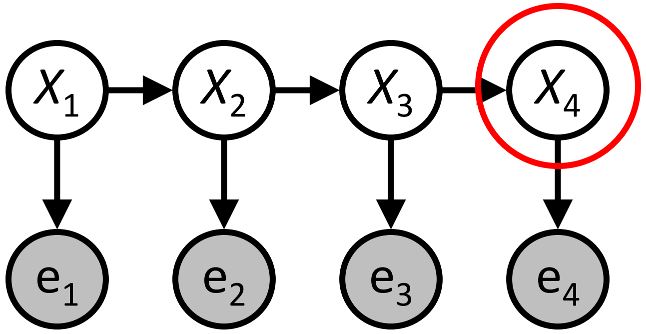
Useful notation:

$X_{a:b} = X_a, X_{a+1}, \ldots, X_b$

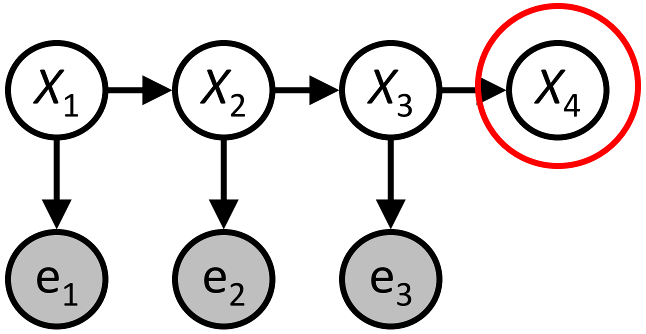
- ***Filtering***: $P(X_t | e_{1:t})$
  - ***belief state***—input to the decision process of a rational agent
- ***Prediction***: $P(X_{t+k} | e_{1:t})$ for $k > 0$
  - evaluation of possible action sequences; like filtering without the evidence
- ***Smoothing***: $P(X_k | e_{1:t})$ for $0 \le k < t$
  - better estimate of past states, essential for learning
- ***Most likely explanation***: $\arg\max_{x_{1:t}} P(x_{1:t} \mid e_{1:t})$
  - speech recognition, decoding with a noisy channel

Inference tasks

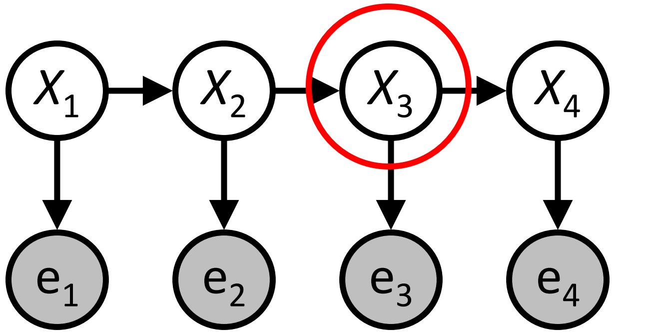
Filtering: $P(X_t|e_{1:t})$



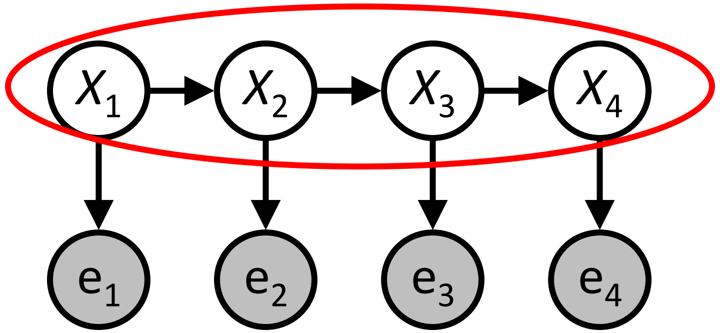
Prediction: $P(X_{t+k}|e_{1:t})$



Smoothing: $P(X_k|e_{1:t})$, $k<t$



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



Filtering / Monitoring

- Filtering, or monitoring, or state estimation, is the task of maintaining the distribution $\mathbf{f}_{1:t} = P(X_t|e_{1:t})$ over time

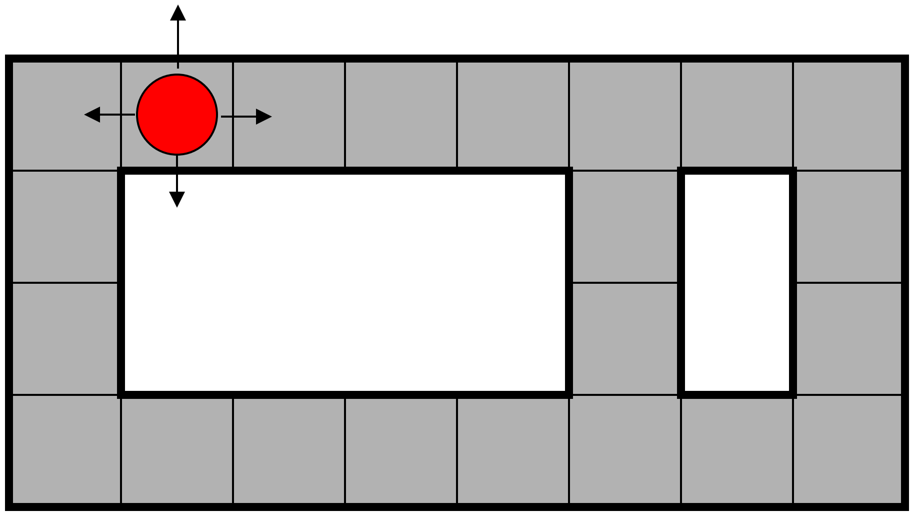
- We start with $\mathbf{f}_0$ in an initial setting, usually uniform

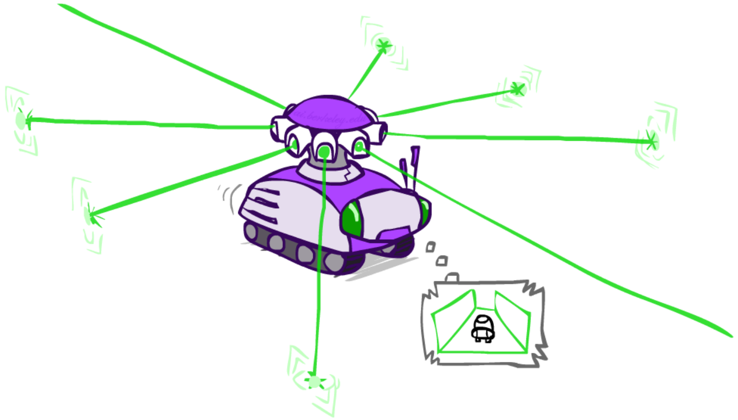
- Filtering is a fundamental task in engineering and science

- The Kalman filter (continuous variables, linear dynamics, Gaussian noise) was invented in 1960 and used for trajectory estimation in the Apollo program; core ideas used by Gauss for planetary observations; ***>1,000,000*** papers on Google Scholar

Example: Robot Localization

Example from
Michael Pfeiffer





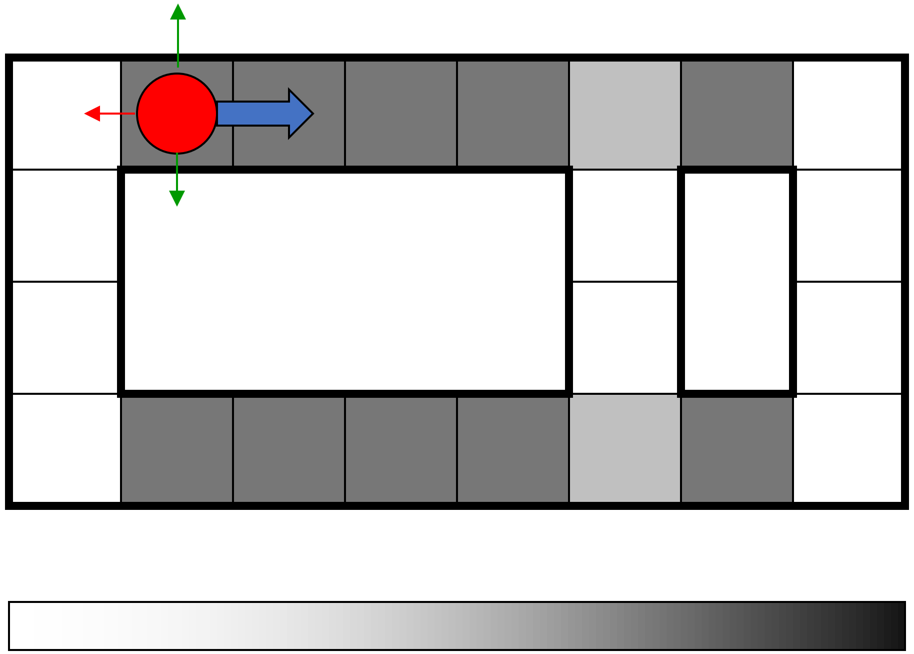
Prob 0 1

t=0

Sensor model: four bits for wall/no-wall in each direction, never more than 1 mistake

Transition model: action may fail with small prob.

Example: Robot Localization

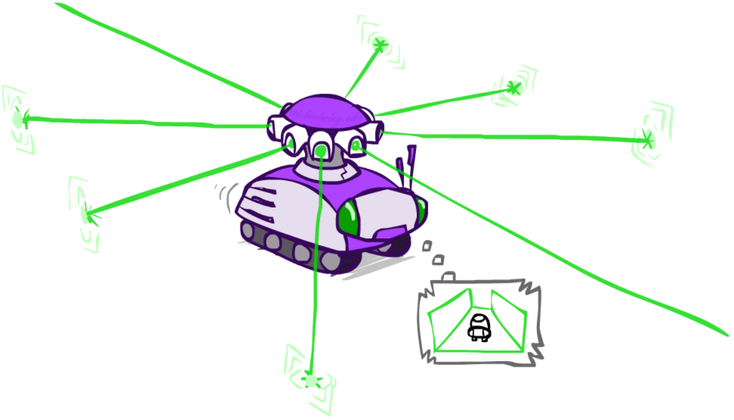


Prob 0 1

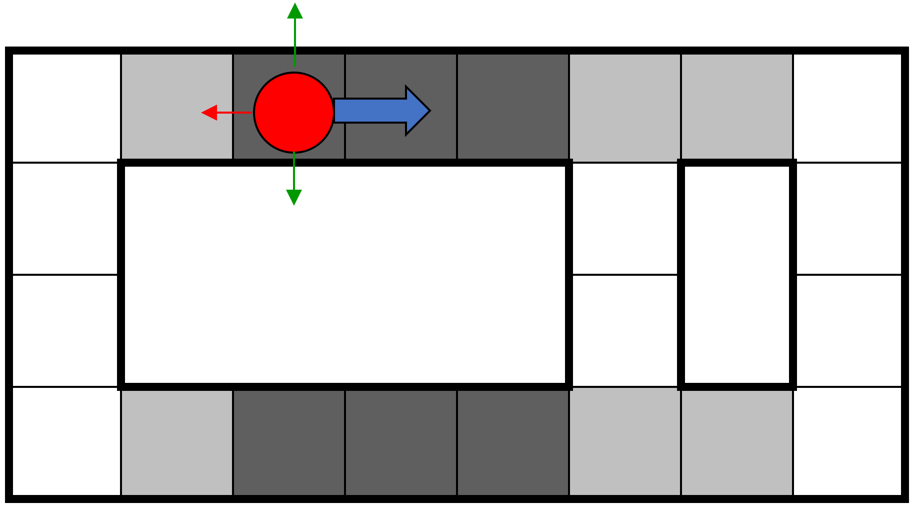
t=1

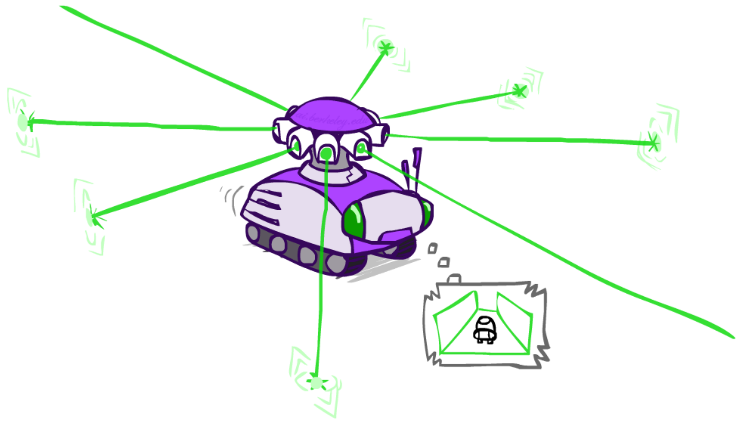
Lighter grey: was ***possible*** to get the reading,

but ***less likely*** (required 1 mistake)



Example: Robot Localization

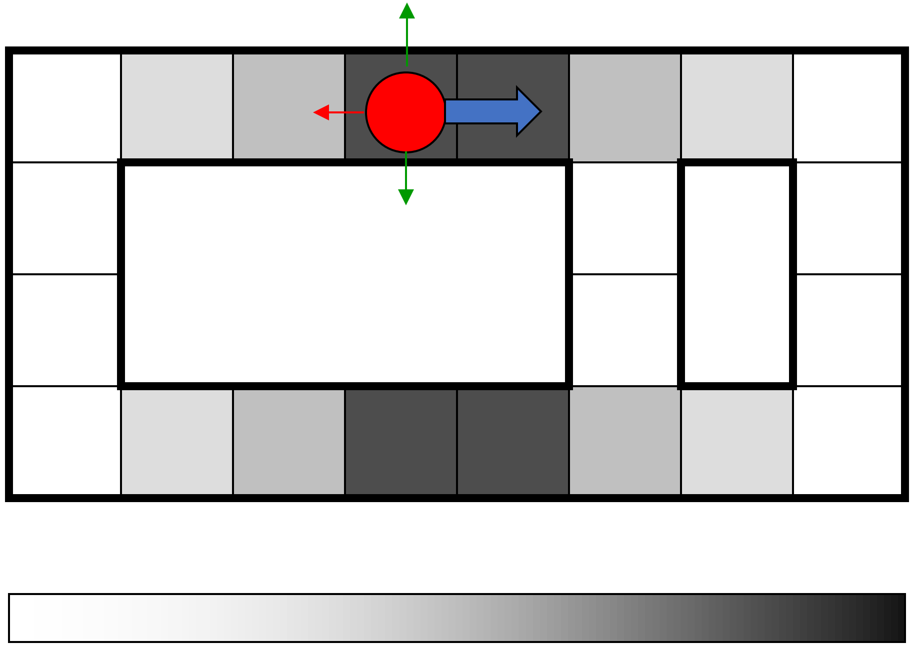




Prob 0 1

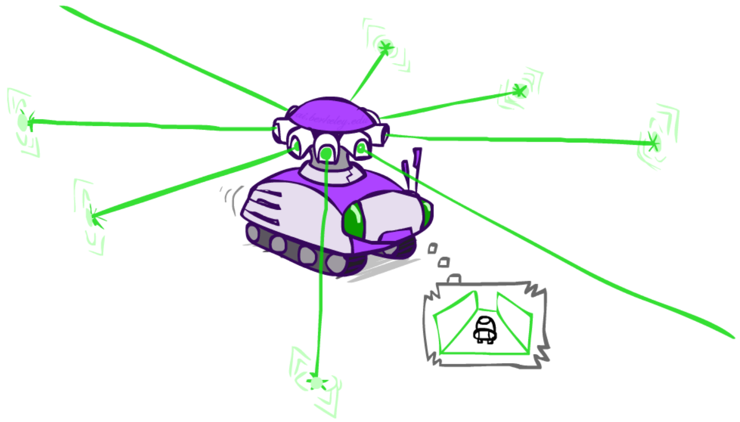
t=2

Example: Robot Localization

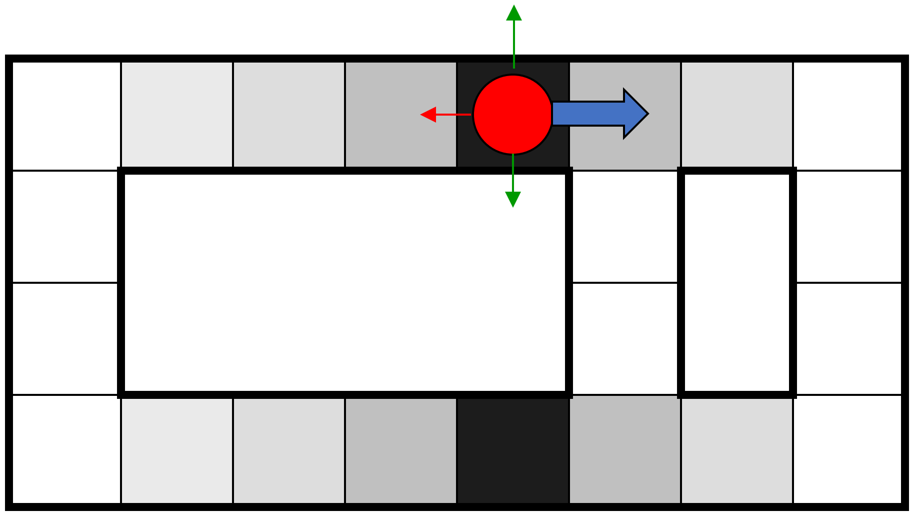


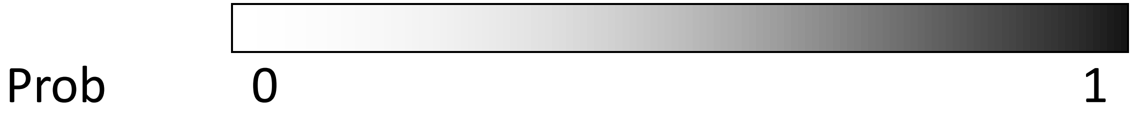
Prob 0 1

t=3

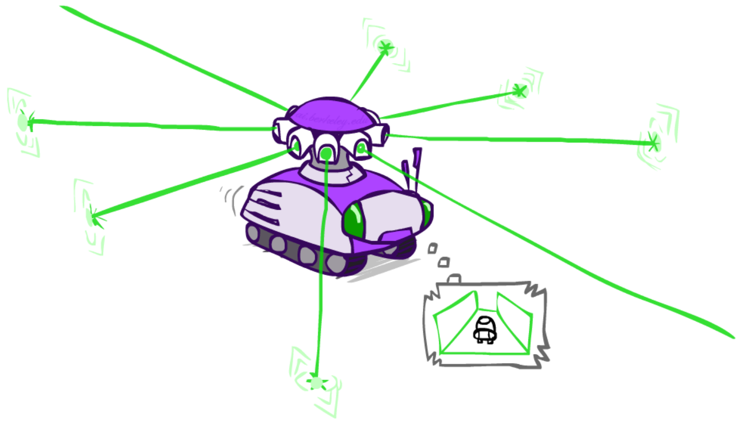


Example: Robot Localization

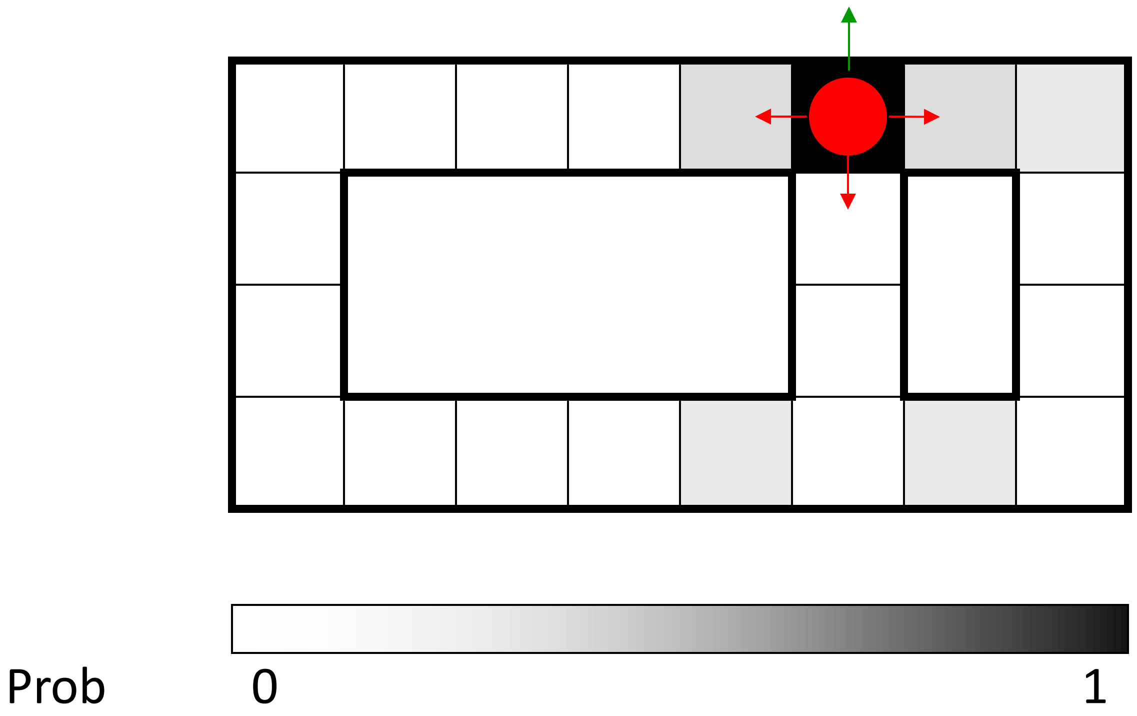




t=4



Example: Robot Localization



t=5