# EECS 445

# Introduction to Machine Learning

## Hidden Markov Models

**Prof. Kutty**

# Announcements

Final exam room assignments published early next week. Students with accommodations: emails will be sent out this week.

Final quiz due this Sunday

Sample exam published Friday; solutions after review session.

DAG

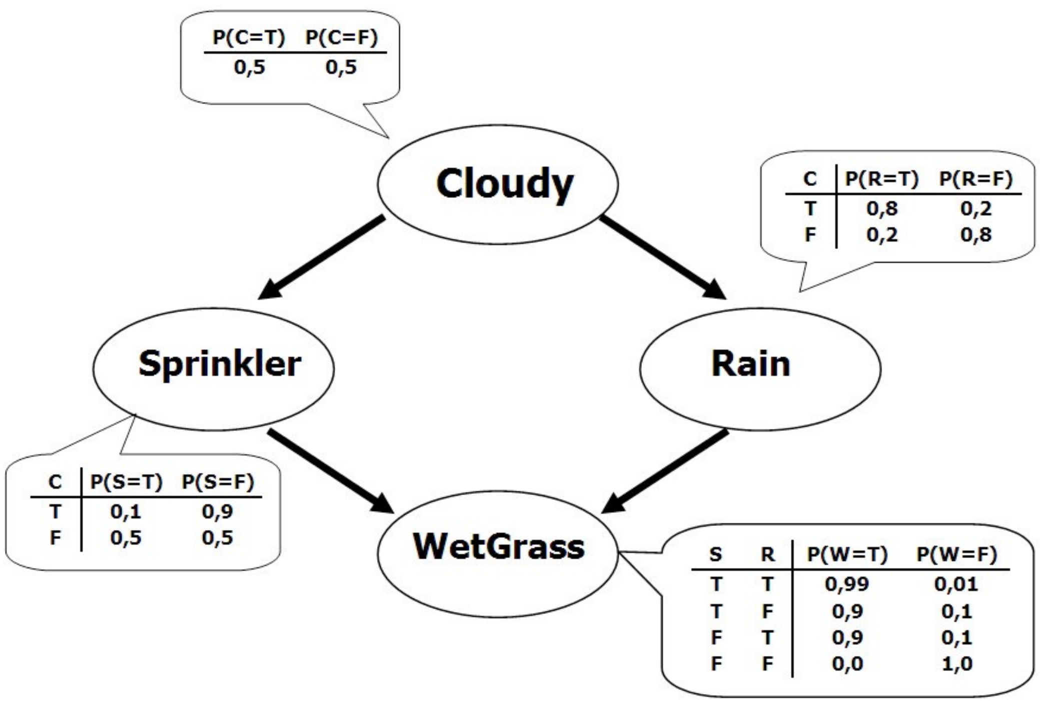
CPT

# Learning Bayesian Networks

# Learning Bayesian Networks

Two Main Problems

1. estimate parameters given graph structure (and data)
2. search over possible graph structures (model sel.)



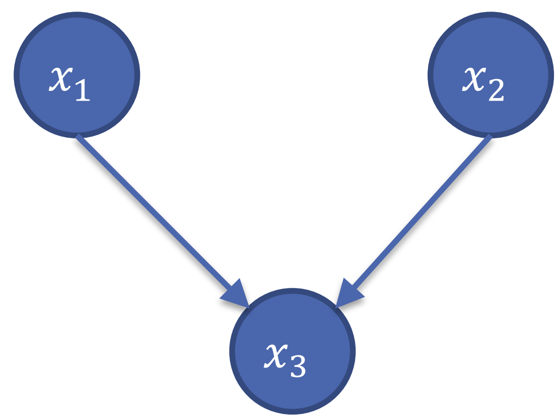
# Learning Bayesian Networks:
# parameter estimation

# Learning Parameters in a Bayesian Network: Setup

- Get a dataset
  - $d = 3$ and each $x_i$ is a binary random variable

| $x_1$ | $x_2$ | $x_3$ |
|---|---|---|
| 0 | 0 | 1 |
| 0 | 1 | 1 |
| 1 | 0 | 0 |
| 1 | 1 | 1 |
| 0 | 1 | 0 |
| 1 | 1 | 1 |
| 0 | 1 | 0 |

- Current guess on relationships between variables
  - we'll see a more systematic approach later



# Parameter Estimation: Example

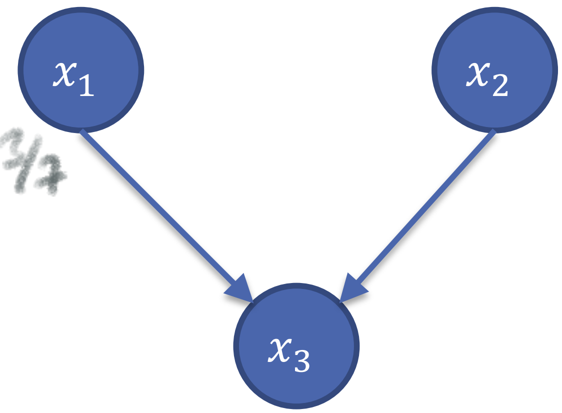
$Pr(x_1, x_2, x_3) = Pr(x_1)\, Pr(x_2)\, Pr(x_3 \mid x_1, x_2)$

$\prod_{i=1}^{n} Pr(x_1, x_2, x_3)$

$\theta_1(x_1)$

| $x_1$ | 0 | 1 |
|---|---|---|
| | ? | ? |

4/7 3/7




$\theta_2(x_2)$

| $x_2$ | 0 | 1 |
|---|---|---|
| | ? | ? |

Dataset

| $x_1$ | $x_2$ | $x_3$ |
|---|---|---|
| 0 | 0 | 1 |
| 0 | 1 | 1 |
| 1 | 0 | 0 |
| 1 | 1 | 1 |
| 0 | 1 | 0 |
| 1 | 1 | 1 |
| 0 | 1 | 0 |

Each of these is a parameter to be estimated

How? Use MLE!

$\theta_3(x_3 | x_1, x_2)$

$x_3$

| $x_1$ | $x_2$ | $\Pr(x_3 = 1 \mid x_1, x_2)$ | $\Pr(x_3 = 0 \mid x_1, x_2)$ |
|---|---|---|---|
| 0 | 0 | ? | ? |
| 0 | 1 | ? 1/3 | ? 2/3 |
| 1 | 0 | ? | ? |
| 1 | 1 | ? | ? |

$Pr(x_3 = 1 \mid x_1 = 0, x_2 = 1) + Pr(x_3 = 0 \mid x_1 = 0, x_2 = 1) = 1$

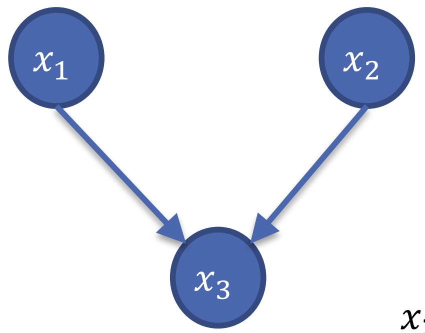
# Parameter Estimation: Example

$d = 3$ and each $x_i$ is a binary random variable

$\theta_1(x_1)$

$x_1$

| 0 | 1 |
|---|---|
| ? | ? |




$\theta_2(x_2)$

$x_2$

| 0 | 1 |
|---|---|
| ? | ? |

$\theta_3(x_3|x_1, x_2)$

$x_3$

| $x_1$ | $x_2$ | $\Pr(x_3 = \mathbf{1}|\boldsymbol{x_1}, \boldsymbol{x_2})$ | $\Pr(x_3 = \mathbf{0}|\boldsymbol{x_1}, \boldsymbol{x_2})$ |
|---|---|---|---|
| 0 | 0 | ? | ? |
| 0 | 1 | ? | ? |
| 1 | 0 | ? | ? |
| 1 | 1 | ? | ? |

| $x_1$ | $x_2$ | $x_3$ |
|---|---|---|
| 0 | 0 | 1 |
| 0 | 1 | 1 |
| 1 | 0 | 0 |
| 1 | 1 | 1 |
| 0 | 1 | 0 |
| 1 | 1 | 1 |
| 0 | 1 | 0 |

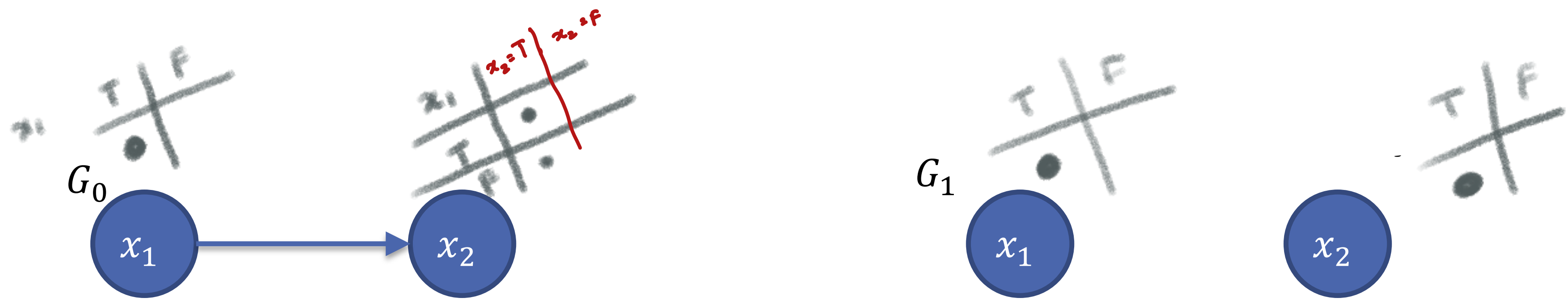
$$\hat{\theta}_i(x_i = v_i | x_{pa_i} = v_{pa_i}) = \frac{n(x_i = v_i, x_{pa_i} = v_{pa_i})}{\sum_{v_i'} n(x_i = v_i', x_{pa_i} = v_{pa_i})}$$

Assume that $n(\cdot)$ returns the count

# Learning Bayesian Networks: learning the graph structure

# Learning Graph Structure

First attempt (doesn't work)

- Idea: Choose graph structure that maximizes log likelihood
- 


$$l(\theta; S_n, G_0) = \sum_{t=1}^{n} \ln \theta_1(x_1^{(t)}) + \ln \theta_2(x_2^{(t)} | x_1^{(t)})$$

$$l(\theta; S_n, G_1) = \sum_{t=1}^{n} \ln \theta_1(x_1^{(t)}) + \ln \theta_2(x_2^{(t)})$$

# Model Selection

## Bayesian Information Criterion (BIC)

number of training data

Log-likelihood

$$BIC(D;\bar{\theta}) = \boxed{l(D;\bar{\theta})} - \boxed{\frac{\#param}{2}} log(n)$$

model complexity

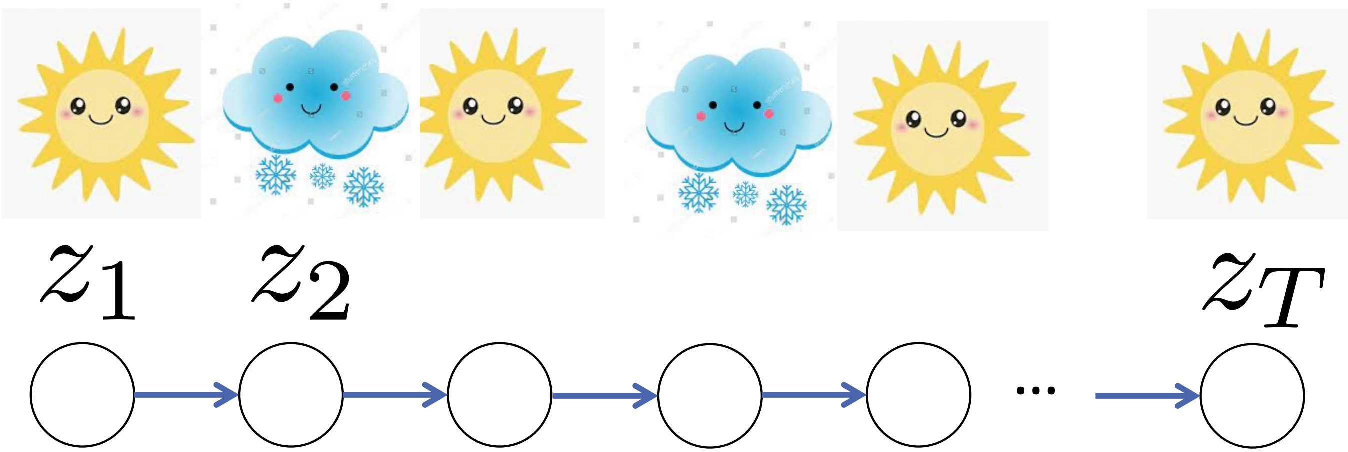
Here we'd want to maximize the BIC.

# Graphical Models:
# Hidden Markov Models

Modeling Sequential Data

# Markov Process: intuition

weather state



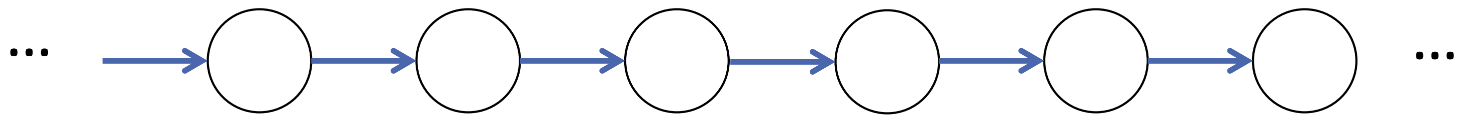

future depends on past via the present

suppose I want to predict tomorrow's weather
if I know today's weather, that is helpful; however
also knowing yesterday's weather will not provide additional information

to see this use d-separation

In determining the next state we don't care what the previous states were
only what the current state is. The other states give no additional information.

# Transition Probabilities




$z_t$ current state $\quad$ $z_{t+1}$ next state

$z_{t+1}$ is chosen according to the probability distribution associated with $z_t$

e.g., suppose $\forall z_t \; z_t \in \{h_1, h_2\}$

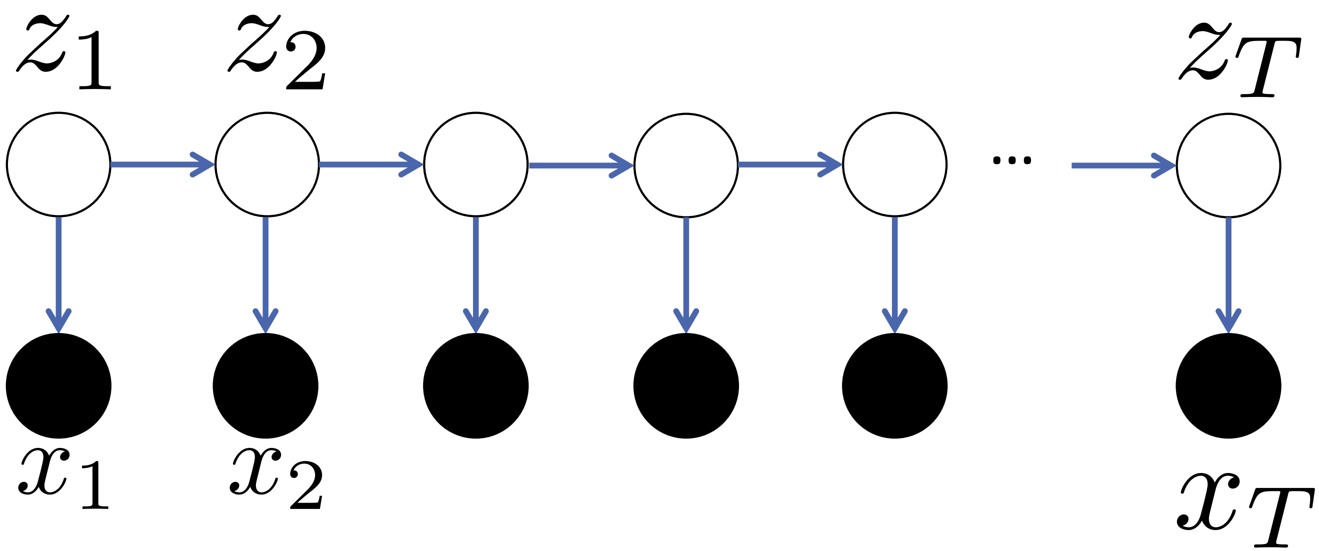
| current state \ next state | $h_1$ | $h_2$ |
|---|---|---|
| $h_1$ | 0.1 | 0.9 |
| $h_2$ | 0.2 | 0.8 |

$P(z_{t+1} = h_1 | z_t = h_1) = 0.1$

# Hidden Markov Model

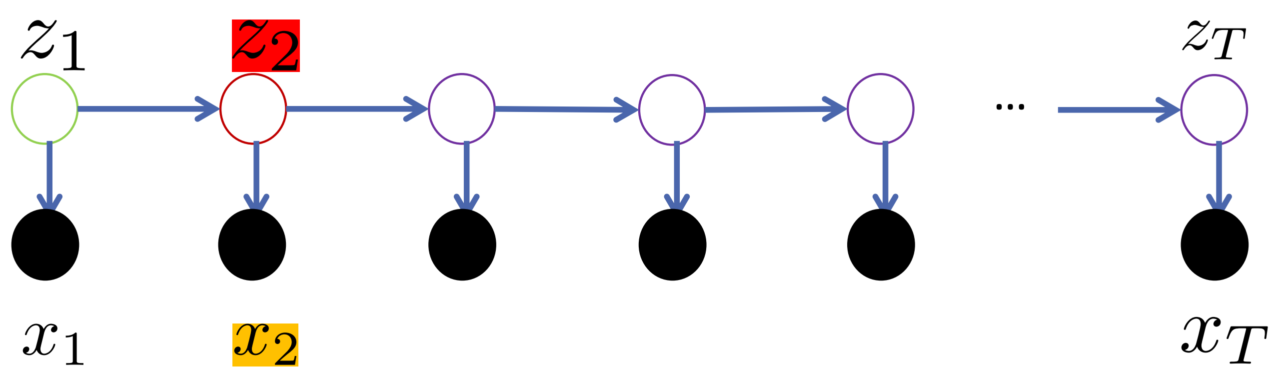
Observed data $X$; Hidden/Latent variables $Z$

**assumption**: the hidden variables are *discrete* random variables



# Hidden Markov Models (intuition)

- encodes two independence properties:
  - Markov process (hidden states): future depends on past via the present
  - current observation independent of all other variables given current hidden state




- **observations** are correlated by hidden states

slide credit: http://ai.berkeley.edu.

*note: this is not an exhaustive list of properties

# Hidden Markov Model
## example application: parts of speech tagging

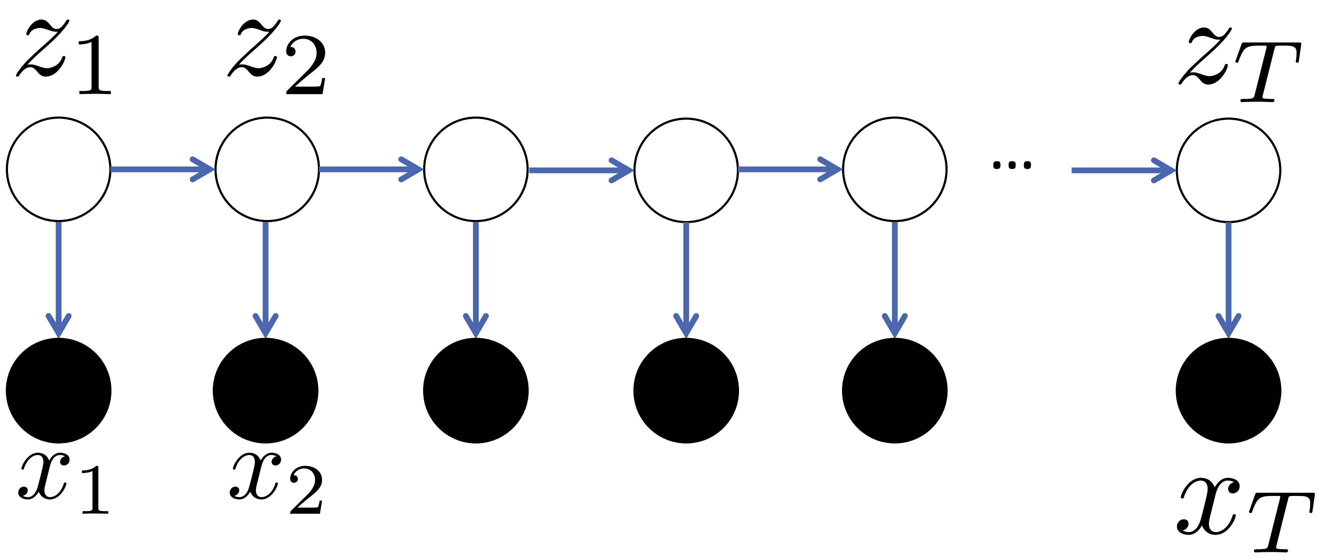
Observed data $X$

**words in a sentence**

Hidden/Latent variables $Z$

**part of speech**

context matters…




| | |
|---|---|
| the output was the mean of the three samples | noun |
| I'm not sure what you mean | verb |
| it's not nice to be mean | adjective |

# Hidden Markov Models

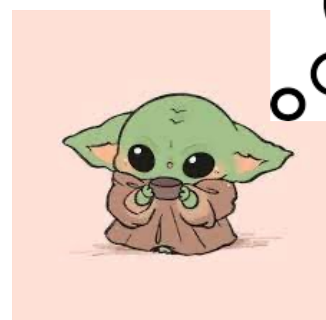
every day is either

hot

or

cold

graduated!

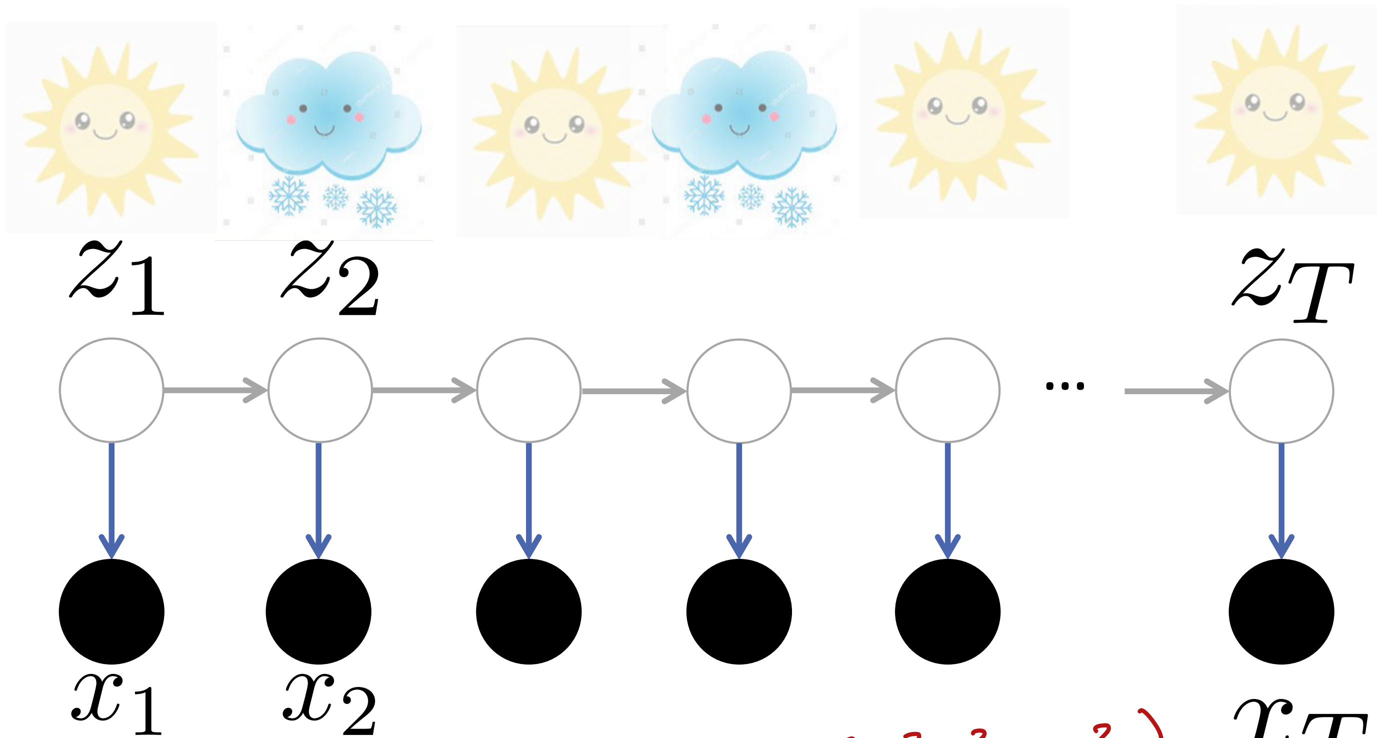
finals season

# ice cream scoops

# Hidden Markov Models: example

E.g., HIDDEN weather state

likelihood function




$$\Pr(x_1, x_2, \dots, x_T, z_1, z_2, z_3 \dots z_T)$$
$$= \Pr(z_1)\ \Pr(x_1|z_1)\ \Pr(x_2|z_2) \dots \Pr(x_T|z_T)\ \Pr(z_2|z_1)\ \Pr(z_3|z_2) \dots \Pr(z_T|z_{T-1})$$

OBSERVED: # ice cream scoops



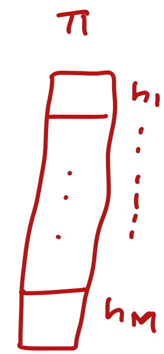
# Example

$H = \{h_1, \dots, h_M\}$

$O = \{o_1, \dots, o_N\}$

Suppose you are given the following model and parameters

- M = 2 H = {hot, cold}
- N = 3 O = {one_scoop, two_scoops, three_scoops}



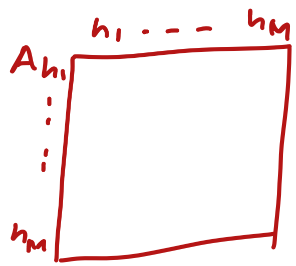

- Assume $\pi(\text{hot}) = 1;\ \ \pi(\text{cold}) = 0$

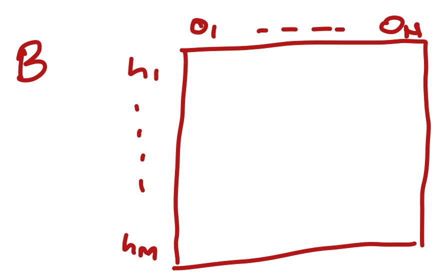
Transition probabilities A:

next state

current state

| | hot | cold |
|---|---|---|
| hot | 0.5 | 0.5 |
| cold | 0.2 | 0.8 |




Emission probabilities B:

observations

current state

| | one_scoop | two_scoops | three_scoops |
|---|---|---|---|
| hot | 0.1 | 0.2 | 0.7 |
| cold | 0.4 | 0.5 | 0.1 |

# Example contd.

Assume $\pi(\text{hot}) = 1; \quad \pi(\text{cold}) = 0$

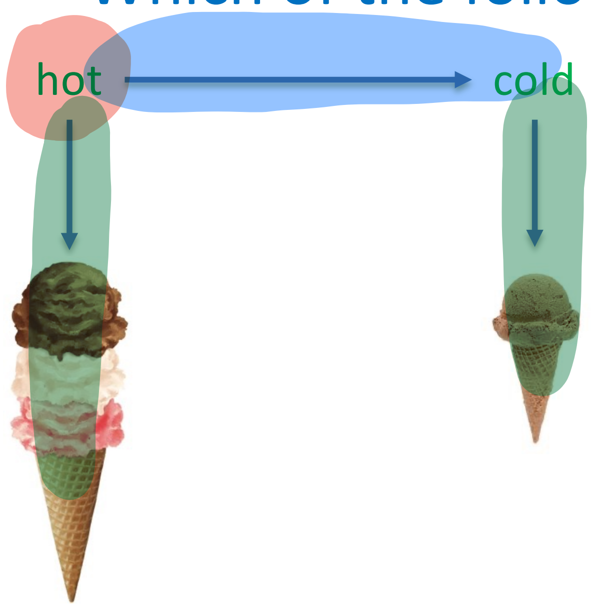
Transition probabilities A:

| current state \ next state | hot | cold |
|---|---|---|
| hot | 0.5 | 0.5 |
| cold | 0.2 | 0.8 |

Emission probabilities B:

| current state \ observations | one_scoop | two_scoops | three_scoops |
|---|---|---|---|
| hot | 0.1 | 0.2 | 0.7 |
| cold | 0.4 | 0.5 | 0.1 |

Which of the following sequences is more likely




# Example contd.

Assume $\pi(\text{hot}) = 1; \quad \pi(\text{cold}) = 0$

Transition probabilities A:

next state

current state

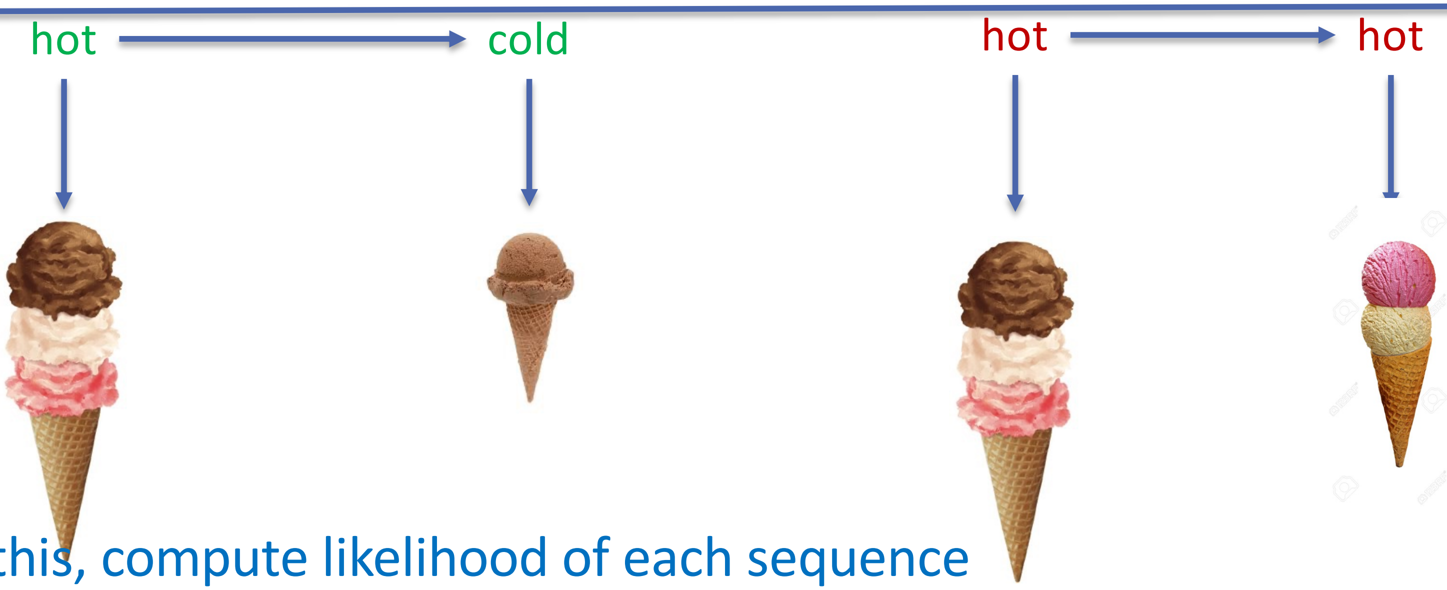
| | hot | cold |
|---|---|---|
| hot | 0.5 | 0.5 |
| cold | 0.2 | 0.8 |

Emission probabilities B:

observations

current state

| | one_scoop | two_scoops | three_scoops |
|---|---|---|---|
| hot | 0.1 | 0.2 | 0.7 |
| cold | 0.4 | 0.5 | 0.1 |




To answer this, compute likelihood of each sequence

$$P(x_1, ..., x_T, z_1, ..., z_T) = \pi(z_1) \prod_{t=1}^{T-1} A(z_t, z_{t+1}) \prod_{t=1}^{T} B(z_t, x_t)$$

# Example: HMM likelihood

Assume $\pi(\text{hot}) = 1; \quad \pi(\text{cold}) = 0$

Transition probabilities A:

next state

| current state | hot | cold |
|---|---|---|
| hot | 0.5 | 0.5 |
| cold | 0.2 | 0.8 |

Emission probabilities B:

observations

| current state | one_scoop | two_scoops | three_scoops |
|---|---|---|---|
| hot | 0.1 | 0.2 | 0.7 |
| cold | 0.4 | 0.5 | 0.1 |

hot → cold


$\pi(\text{hot}) \times A(\text{hot}, \text{cold}) \times B(\text{hot}, 3) \times B(\text{cold}, 1)$

is more likely than

hot → hot


$\pi(\text{hot}) \times A(\text{hot}, \text{hot}) \times B(\text{hot}, 3) \times B(\text{hot}, 2)$

# Hidden Markov Models

Transition Probabilities

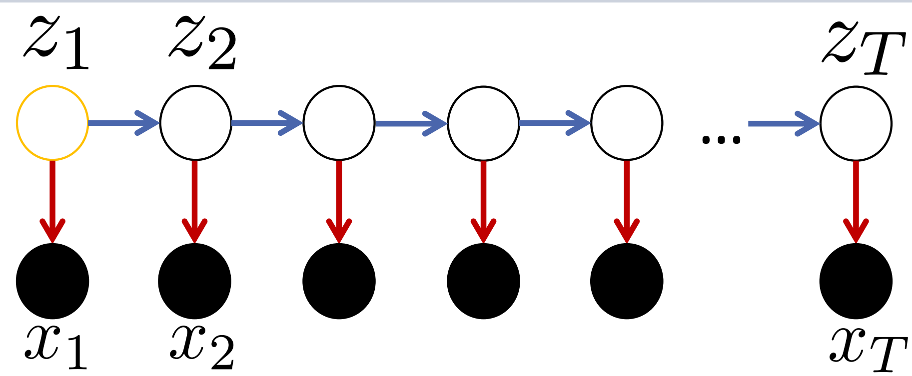
$$A(h_i, h_j) = P(z_{t+1} = h_j | z_t = h_i)$$

Emission Probabilities

$$B(h_i, o_l) = P(x_t = o_l | z_t = h_i)$$

Starting State Prob.

$$\pi(h_i) = P(z_1 = h_i)$$

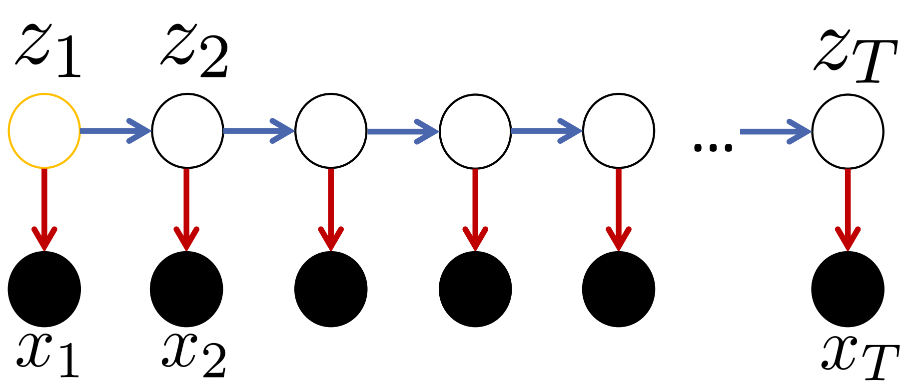



Likelihood of a given sequence

$$P(x_1, \ldots, x_T, z_1, \ldots, z_T) = \underbrace{\pi(z_1)}_{\text{starting state}} \overbrace{\prod_{t=1}^{T-1} A(z_t, z_{t+1})}^{\text{transitions}} \underbrace{\prod_{t=1}^{T} B(z_t, x_t)}_{\text{emissions}}$$

When you do not have access to *hidden* states…

We aim to infer it *given observations*.



# Decoding HMM

In general, we are not given the underlying sequence of states.

We aim to infer it *given observations*.

e.g.,

Given $\pi(\text{hot}) = 1; \quad \pi(\text{cold}) = 0$

Transition probabilities A:

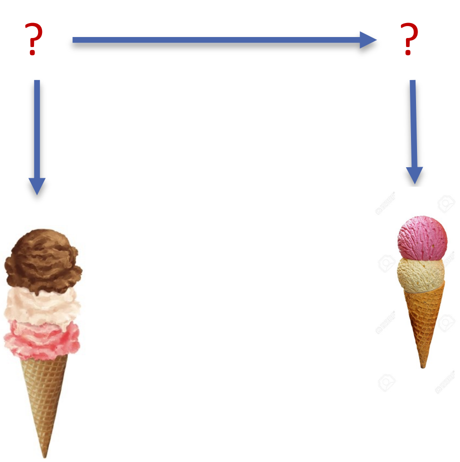
next state

| current state | | hot | cold |
|---|---|---|---|
| | hot | 0.5 | 0.5 |
| | cold | 0.2 | 0.8 |

Emission probabilities B:

observations

| | one_scoop | two_scoops | three_scoops |
|---|---|---|---|
| hot | 0.1 | 0.2 | 0.7 |
| cold | 0.4 | 0.5 | 0.1 |



# Decoding HMM (inefficiently)

e.g.,

Given $\pi(\text{hot}) = 1; \quad \pi(\text{cold}) = 0$

Transition probabilities A:

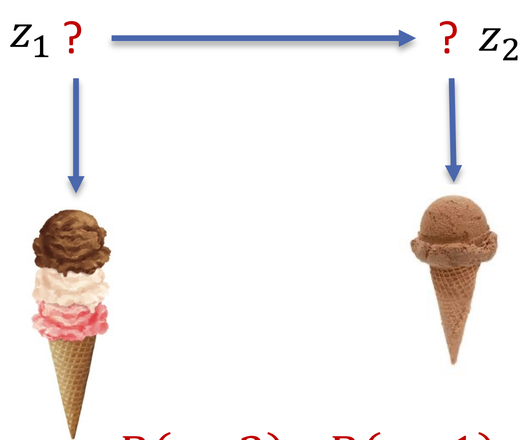
next state

| current state | hot | cold |
|---|---|---|
| hot | 0.5 | 0.5 |
| cold | 0.2 | 0.8 |

Emission probabilities B:

observations

| | one_scoop | two_scoops | three_scoops |
|---|---|---|---|
| hot | 0.1 | 0.2 | 0.7 |
| cold | 0.4 | 0.5 | 0.1 |




$$\arg\max_{z_1, z_2 \in \{\text{hot,cold}\}} \pi(z_1) \times A(z_1, z_2) \times B(z_1, 3) \times B(z_2, 1) = (\text{hot}, \text{cold})$$

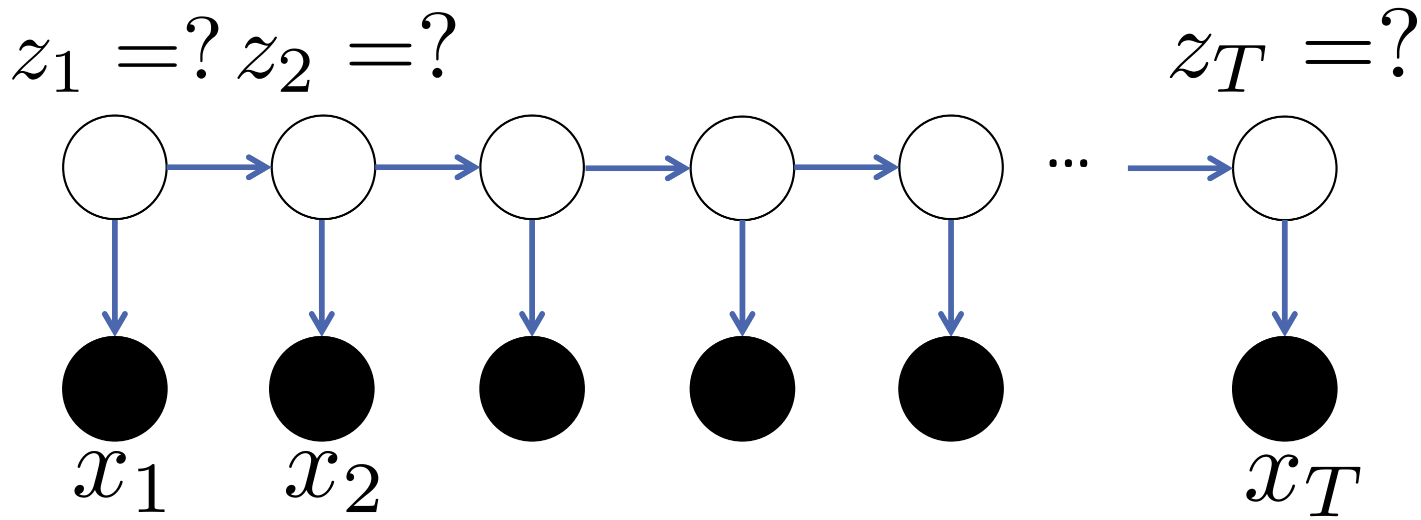
$$\pi(z_1 = \text{hot}) \times A(z_1 = \text{hot}, z_2 = \text{hot}) \times B(z_1 = \text{hot}, 3) \times B(z_2 = \text{hot}, 1) = 0.5 \times 0.7 \times 0.1$$

$$\pi(z_1 = \text{hot}) \times A(z_1 = \text{hot}, z_2 = \text{cold}) \times B(z_1 = \text{hot}, 3) \times B(z_2 = \text{cold}, 1) = 0.5 \times 0.7 \times 0.4$$

$$\pi(z_1 = \text{cold}) \times A(z_1 = \text{cold}, z_2 = \text{hot}) \times B(z_1 = \text{cold}, 3) \times B(z_2 = \text{hot}, 1) = 0$$

$$\pi(z_1 = \text{cold}) \times A(z_1 = \text{cold}, z_2 = \text{cold}) \times B(z_1 = \text{cold}, 3) \times B(z_2 = \text{cold}, 1) = 0$$

# Decoding HMM




**Given**: the observations and the model parameters

**Goal:** infer the underlying hidden states

$$\underset{z_1,\ldots,z_T}{\operatorname{argmax}} P(x_1, \ldots, x_T, z_1, \ldots, z_T; \theta)$$

$$= \underset{z_1,\ldots,z_T}{\operatorname{argmax}} \pi(z_1) \prod_{t=1}^{T-1} A(z_t, z_{t+1}) \prod_{t=1}^{T} B(z_t, x_t)$$

# Decoding HMM

**Given**: the observations and the model parameters

**Goal:** infer the underlying hidden states

$$\underset{z_1,\ldots,z_T}{\operatorname{argmax}} P(x_1,\ldots,x_T,z_1,\ldots,z_T;\theta)$$

**First attempt:**

enumerate all possibilities and choose

**Bad idea in general**. Why?

# Decoding HMM (efficiently):
# Viterbi Algorithm