COMS 4771
Probabilistic Reasoning via Graphical Models

Nakul Verma

Last time...

- Dimensionality Reduction
  - Linear vs non-linear Dimensionality Reduction
- Principal Component Analysis (PCA)
- Non-linear methods for doing dimensionality reduction

Graphical Models

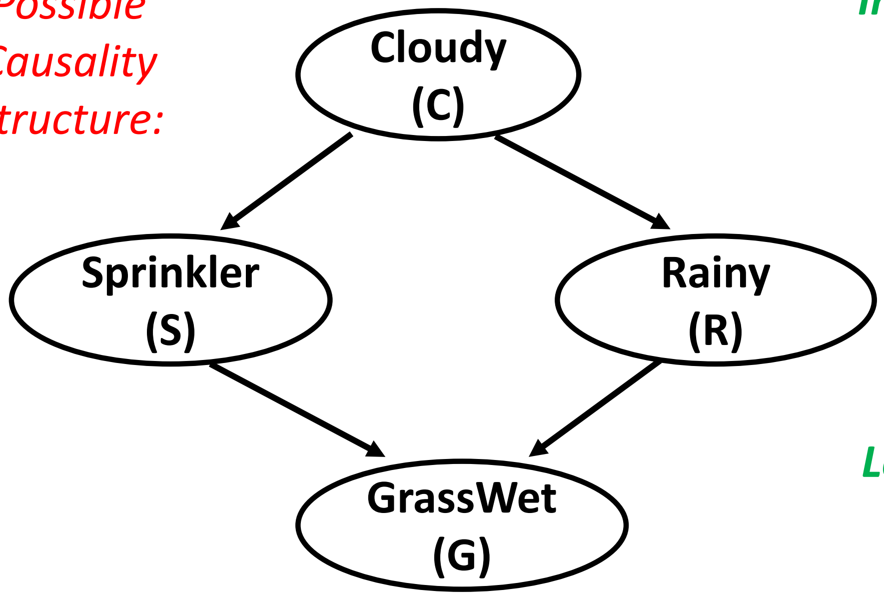
A probabilistic model where a graph represents the conditional dependence structure among the variables.

Provides a compact representation of the joint distribution!

Example:

Four variables of interest – cloudiness, raining, sprinkler, grass_wet

Possible Causality Structure:



Inference questions:

- *What is the probability it rained given the grass is wet?*
- *What is the chance that the sprinkler was off given grass is wet and it is not cloudy?*

Learning questions:

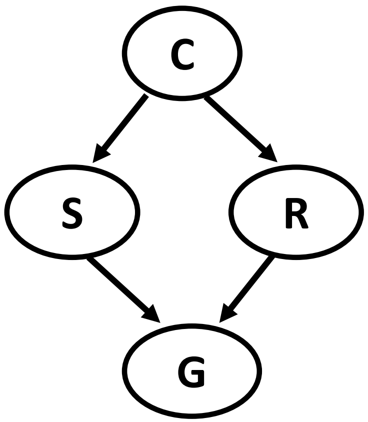
- *What is the most likely GM structure and connection weights that models the data?*

Graphical Models: Representation

There are two kinds of Graphical Models

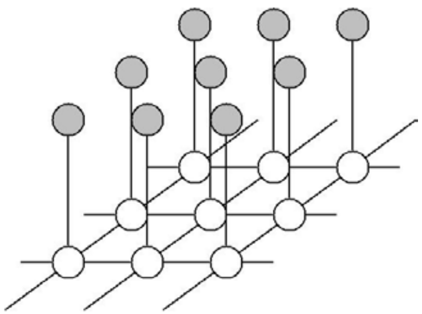
Directed models – Bayesian Networks

Edge direction typically denotes potential causality



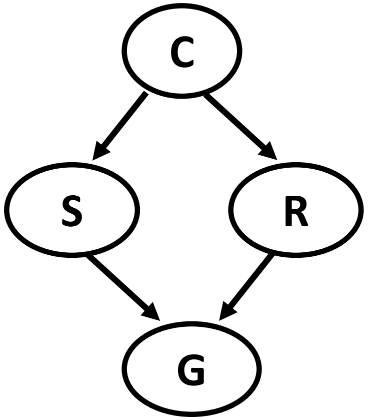
Undirected models – Markov Random Fields (MRFs)

Edge connection typically denotes potential co-occurrence



Bayesian Networks

What is the joint probability for these variables?



$$P(C, S, R, G)$$

$$= P(C)P(R|C)P(S|R,C)P(G|S,R,C)$$

Chain rule

$$= P(C)P(R|C)P(S|C)P(G|S,R)$$

due to the (in)dependencies asserted by the parent-child relationships

In general:

$$P(X_1, \ldots, X_d) \quad = \prod_{i=1}^{d} P(X_i \mid \mathrm{parent}(X_i))$$

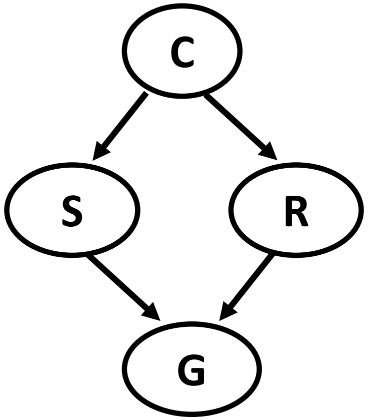
That is: a variable is independent of its ancestors given the parents.

Bayesian Networks: Inference

$$P(C,S,R,G) = P(C)P(R|C)P(S|C)P(G|S,R)$$

P(C=1)
0.5

C	P(S=1|C)
0	0.5
1	0.1

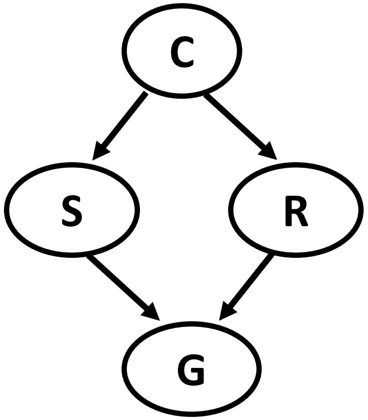


C	P(R=1|C)
0	0.2
1	0.8

S	R	P(G=1|S,R)
0	0	0.0
0	1	0.9
1	0	0.9
1	1	0.99

*These conditional probability tables (CPT) are enough to **completely** specify the joint distribution!*

Bayesian Networks: Inference



$$P(C,S,R,G) = P(C)P(R|C)P(S|C)P(G|S,R)$$

Q: What is the probability of sprinkler being on given the grass is wet?

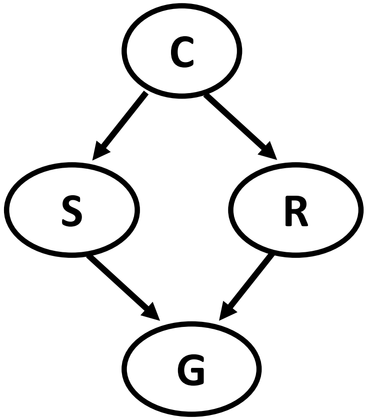
$$P(S=1|G=1) = \frac{P(S=1,G=1)}{P(G=1)} = \frac{0.2781}{0.6471} = 0.430$$

$$\begin{aligned} P(G=1) &= \sum_{c,s,r} P(C=c,S=s,R=r,G=1) \\ &= \sum_{c,s,r} P(C=c)P(R=r|C=c)P(S=s|C=c)P(G=1|S=s,R=r) \\ &= 0.6471 \end{aligned}$$

$$P(S=1,G=1) = \sum_{c,r} P(C=c,S=1,R=r,G=1) = \cdots = 0.2781$$

Bayesian Networks: Learning Parameters

$$P(C,S,R,G) = P(C)P(R|C)P(S|C)P(G|S,R)$$



Learning the parameters knowing the structure

ie, estimate the CPTs from observations

Simply do the likelihood estimates (ie, counts)

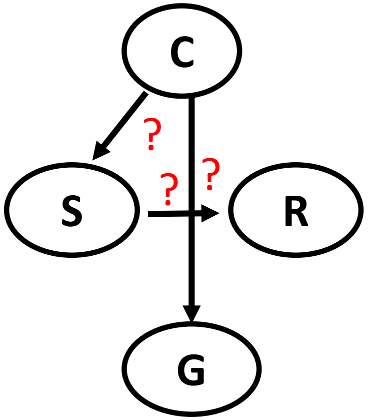
$$\hat{P}_{\mathrm{ML}}(G=g|S=s,R=r) = \frac{\#(G=g,S=s,R=r)}{\#(S=s,R=r)}$$

etc …

Issue: assigns zero prob. for unseen combinations in data. How to fix that?

Bayesian Networks: Learning Structure

$$P(C,S,R,G) = P(C)P(R|C)P(S|R,C)P(G|S,R,C)$$



Learning the unknown structure between the variables

General

- Test of conditional independencies in data
- Grow-Shrink Markov Blanket algorithm

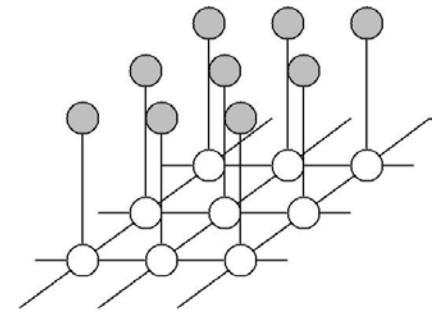
Assumed structure:

- Tree structure: Chow-Liu algorithm
- Small cliques: variations on Chow-Liu

NP-hard to find the optimal structure

Markov Random Fields (MRFs)

Graphical models with undirected connections

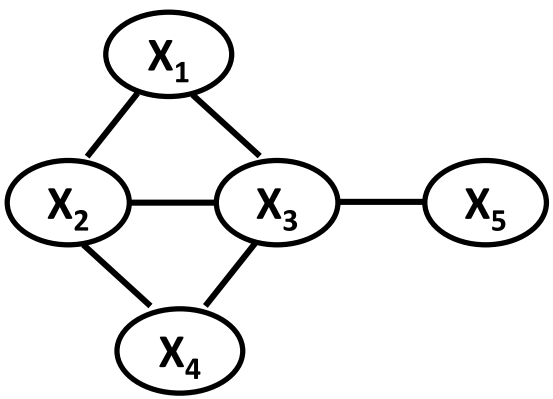


$$P(X_1, \ldots, X_d) = \frac{1}{Z} \prod_{C \in \text{max-cliques}(G)} \phi_C(X_C)$$

normalizer (so things integrate to 1), aka the partition function

Clique potentials, typically the relative frequency of variable co-occurrence in a clique

Example: five variable graph



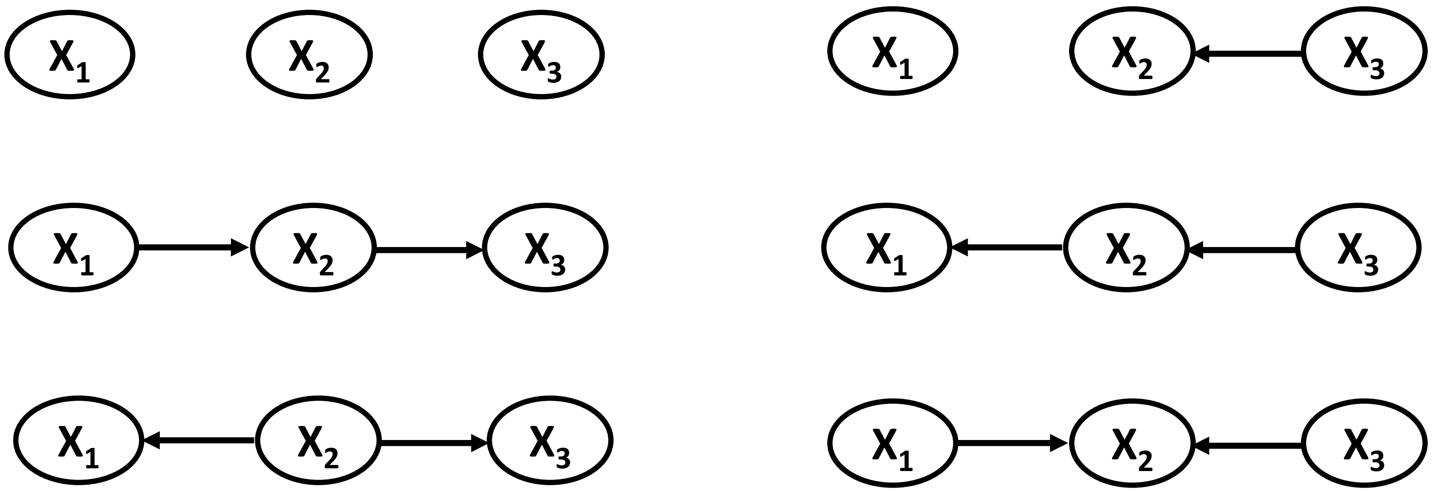
What are the max-cliques?

$$P(X_1, \ldots, X_5) \propto \phi_1((X_1, X_2, X_3))\ \phi_2((X_2, X_3, X_4))\ \phi_3((X_3, X_5))$$

A Closer Look at (In)dependencies in GMs

What are the (conditional) independencies asserted by the following graphical models?

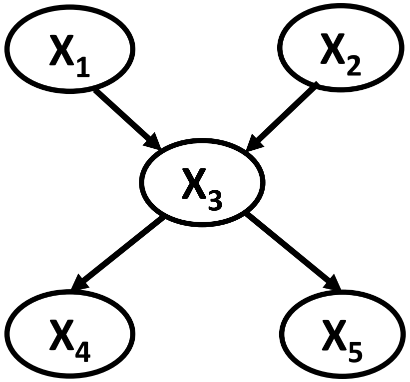
(directed)



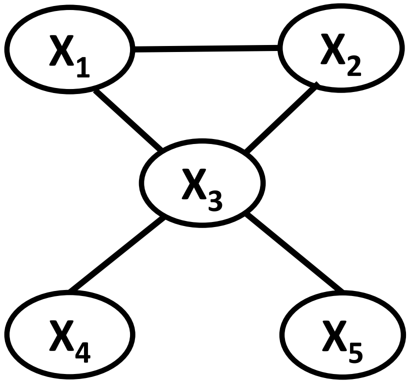
(undirected)

Relation Between Directed & Undirected GM

What are the (conditional) independencies asserted by the following directed model?



What is the equivalent undirected model?



GM Special Case: Time Series Model

A time series model:

A family of distributions over a sequence of random variables X_1, X_2,… that is indexed by a totally ordered indexing set (often referred to as *time*)

… $\mathbf{X}_{t-2}$ $\mathbf{X}_{t-1}$ $\mathbf{X}_{t}$ $\mathbf{X}_{t+1}$ $\mathbf{X}_{t+2}$ …

past states ***current state*** *future states*

Many applications:

- Financial/Economic data over time
- Climate data
- Speech and natural language
- …

Markov Models

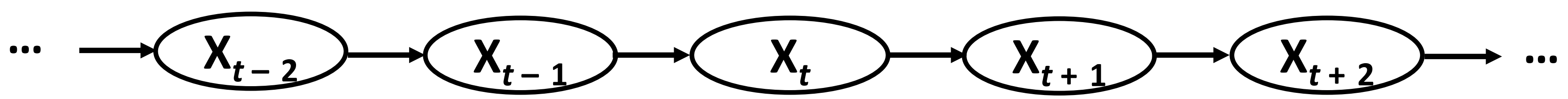
Markov Model:

A time series model with the property:

The conditional distribution of the next state X_{t+1} given all the previous states X_i ($i \leq t$) only depends on the current state X_t

$$P(X_{t+1} \mid X_t, X_{t-1}, X_{t-2}, \ldots) = P(X_{t+1} \mid X_t)$$

The corresponding graphical model:



also known as
a Markov chain

Markov Chains: Distributions

To specify a Markov Chain:

Need to specify the distribution of the initial state: X_1

Need to specify the conditional distribution: X_{t+1} given X_t

This is often called the transition matrix

(*We will focus on finite size state space, say, d different states*)

Initial state distribution:

$$P(X_1 = i) = \pi_i$$

Conditional distribution:

$$P(X_{t+1} = j \mid X_t = i) = A_{ij}$$

can be summarized in a d x d matrix A

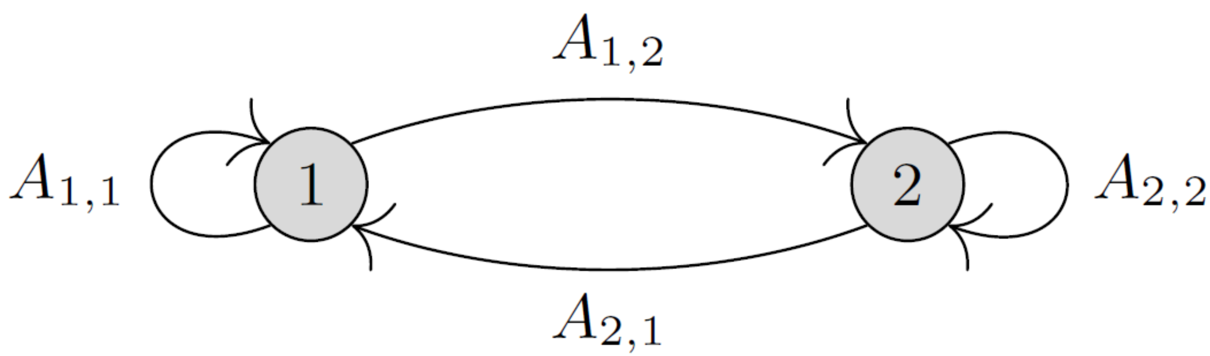
A is row stochastic

Markov Chain: Example

State space: {1,2}

Parameters:

$$\boldsymbol{\pi} = \begin{array}{c} \text{state 1} \\ \text{state 2} \end{array} \begin{pmatrix} 0.1 \\ 0.9 \end{pmatrix}, \quad \boldsymbol{A} = \begin{array}{cc} & \begin{array}{cc} \text{state 1} & \text{state 2} \end{array} \\ \begin{array}{c} \text{state 1} \\ \text{state 2} \end{array} & \begin{pmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \end{pmatrix} \end{array}.$$

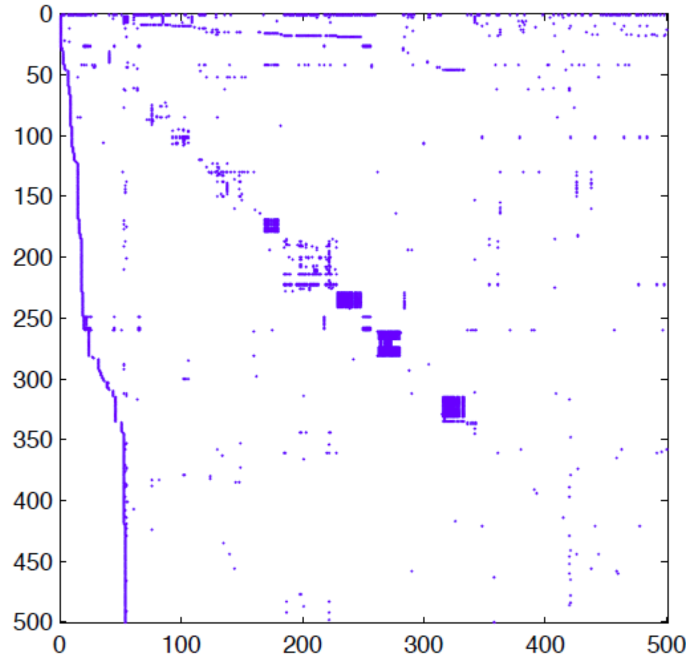


What is the probability of seeing the random sequence: 2,2,2,1,1,2,2,1 ?

$$\pi_2 \cdot A_{2,2} \cdot A_{2,2} \cdot A_{2,1} \cdot A_{1,1} \cdot A_{1,2} \cdot A_{2,2} \cdot A_{2,1} \approx 0.004355$$

Markov Chain: Example - PageRank

Web graph: vertices – webpages, edges – links between webpages



link structure for 500 webpages

Question: how popular is a given webpage *i* ?

Possible answer:

proportional to the probability that a random walk ends on page *i*.

$$P(X_t = i)$$

(for some large t)

Markov Chain: Marginals

Let's calculate the following probabilities:

$$P(X_1 = i) \quad = \pi_i$$

$$\begin{aligned}
P(X_2 = i) \quad &= \sum_j P(X_1 = j, X_2 = i) \\
&= \sum_j P(X_1 = j) \cdot P(X_2 = i \mid X_1 = j) \\
&= \sum_j \pi_j A_{j,i} \\
&= i^{\text{th}} \text{ entry of } \pi^\top A \quad = \left(\pi^\top A\right)_i
\end{aligned}$$

$$P(X_3 = i) \quad = \ldots = \left(\pi^\top A A\right)_i$$

$$P(X_t = i) \quad = \left(\pi^\top A^{t-1}\right)_i$$

for the PageRank example, does this converge to a stable value for large t?

Markov Chain: Limiting Behavior

$$P(X_t = i) = \left(\pi^\top A^{t-1}\right)_i$$

Question does/can P(X_t) have a limiting behavior?

Equivalent to asking:

does $\lim_{t\to\infty} A^t$ approach a limiting matrix $\begin{pmatrix} \text{--- q ---} \\ \text{--- q ---} \\ \text{--- q ---} \end{pmatrix}$ (with identical rows) ?

(a sufficient condition)

For such an *A*, it must satisfy:

$$\lim_{t\to\infty} A^t = \left(\lim_{t\to\infty} A^{t-1}\right) A = \begin{pmatrix} \text{--- q ---} \\ \text{--- q ---} \\ \text{--- q ---} \end{pmatrix} A = \begin{pmatrix} \text{--- q ---} \\ \text{--- q ---} \\ \text{--- q ---} \end{pmatrix}$$

Equivalently:

$$qA = q$$

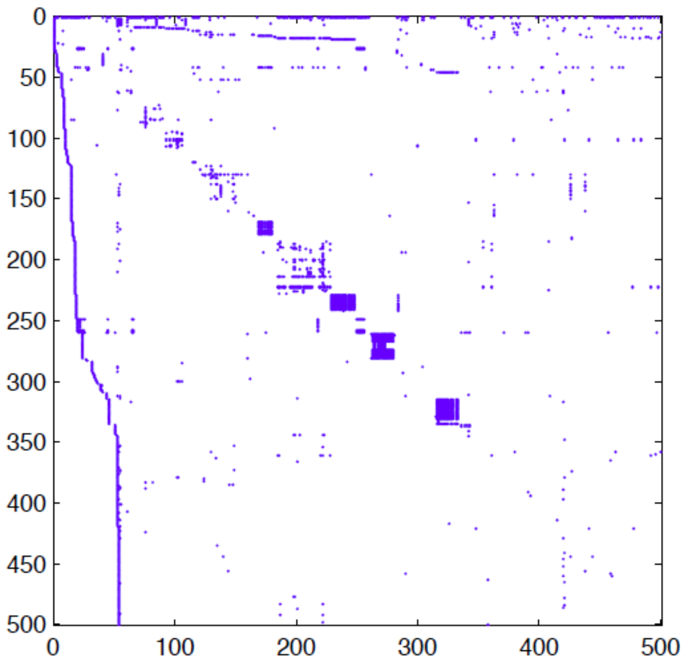
ie, q is the ***left*** *eigenvector of A with eigenvalue 1!*

q unique whenever there is no multiplicity of eigenvalue 1

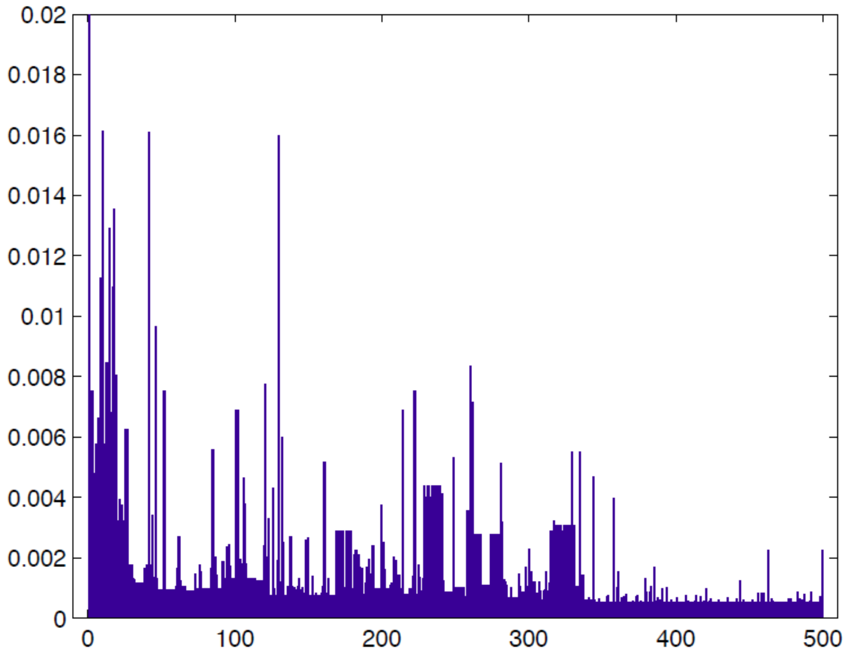
such a q is called the stationary distribution of A

PageRank Example

Web graph doesn't have a unique stationary distribution, but can add some regularity to the link matrix A. That is $\tilde{A} = A + \varepsilon 1$



link structure for 500 webpages



(regularized stationary distribution)

Popularity of a given webpage i is proportional to the i^{th} component of the (regularized) stationary distribution

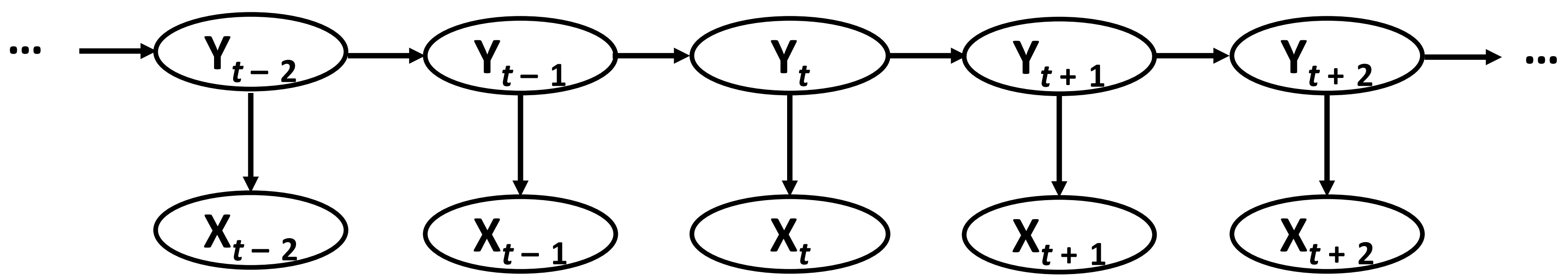
Markov Models with Unobserved Variable

Hidden Markov Model (HMM): A Markov chain on $\{(X_t, Y_t)\}_t$

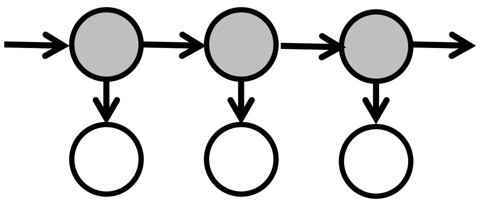
Some properties:

- Y_t is unobserved / hidden variable; only X_t is observed.
- Conditioned on Y_t, X_t is independent of all other variables!

The corresponding graphical model:



Hidden Markov Models (HMMs) Applications



Natural Language Processing

- Observed: words in a sentence
- Unobserved: words' part-of-speech or other word semantics

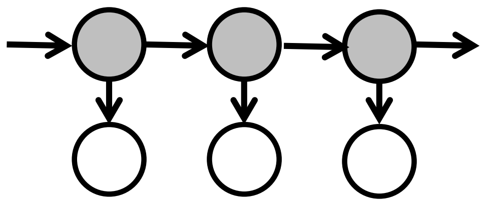
Bioinformatics

- Observed: Amino acids in a protein
- Unobserved: indicators of evolutionary conservation

Speech Recognition

- Observed: Recorded speech
- Unobserved: The phonemes the speaker intended to vocalize

HHMs Parameters



We will focus on discrete state space:

X_t takes values { 1, …, *D* } (observed)

Y_t takes values { 1, …, *K* } (hidden)

We need the initial state distribution on Y_1

$$P(Y_1 = i) = \pi_i$$

Need to specify a *K* x *K* transition matrix *A* from Y_t to Y_{t+1}

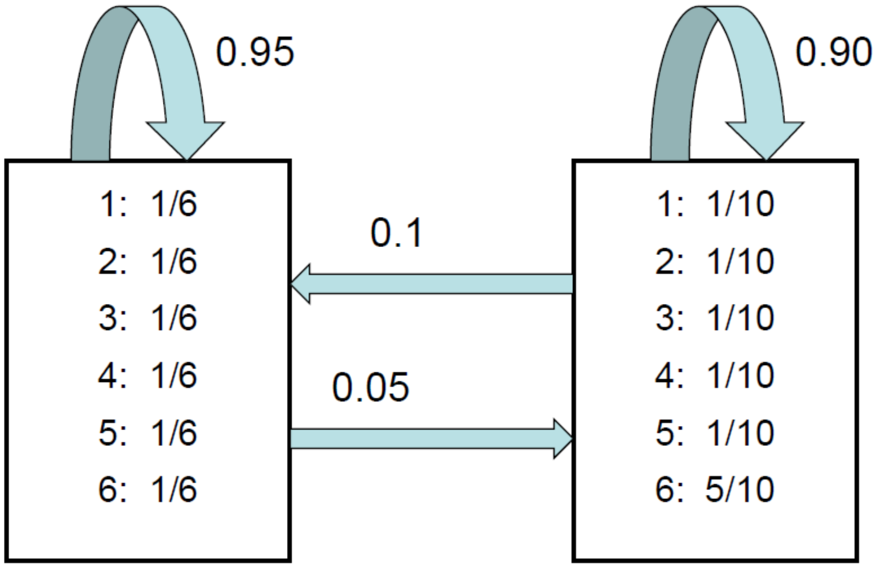
$$P(Y_{t+1} = j \mid Y_t = i) = A_{ij}$$

Need to specify a *K* x *D* emission matrix B from Y_t to X_t

$$P(X_t = j \mid Y_t = i) = B_{ij}$$

Both A and B are row stochastic

HHM: Example – Dishonest Casino



Casino die-rolling game:

Randomly switch between two possible dice: one is fair and one is loaded.

HMM Parameters

$$\boldsymbol{A} = \begin{array}{c c} & \begin{array}{cc} \text{fair die} & \text{loaded die} \end{array} \\ \begin{array}{c} \text{fair die} \\ \text{loaded die} \end{array} & \begin{pmatrix} 0.95 & 0.05 \\ 0.10 & 0.90 \end{pmatrix} \end{array}, \quad \boldsymbol{B} = \begin{array}{c c} & \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{array}{c} \text{fair die} \\ \text{loaded die} \end{array} & \begin{pmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{2} \end{pmatrix} \end{array},$$

$\pi = (1,0)$ [the casino starts off with the fair die]

Problem: based on the sequence of rolls, guess which die was used at each time

HHM Learning and Inference Problems

Conditional Probabilities (filtering/smoothing)

- Given: parameters $\theta = (\pi, A, B)$, and the observation $X_{1:T}$
- Goal: What is the conditional probability of $Y_{1:T}$?

$$P(Y_{1:T} \mid X_{1:T}, \theta)$$

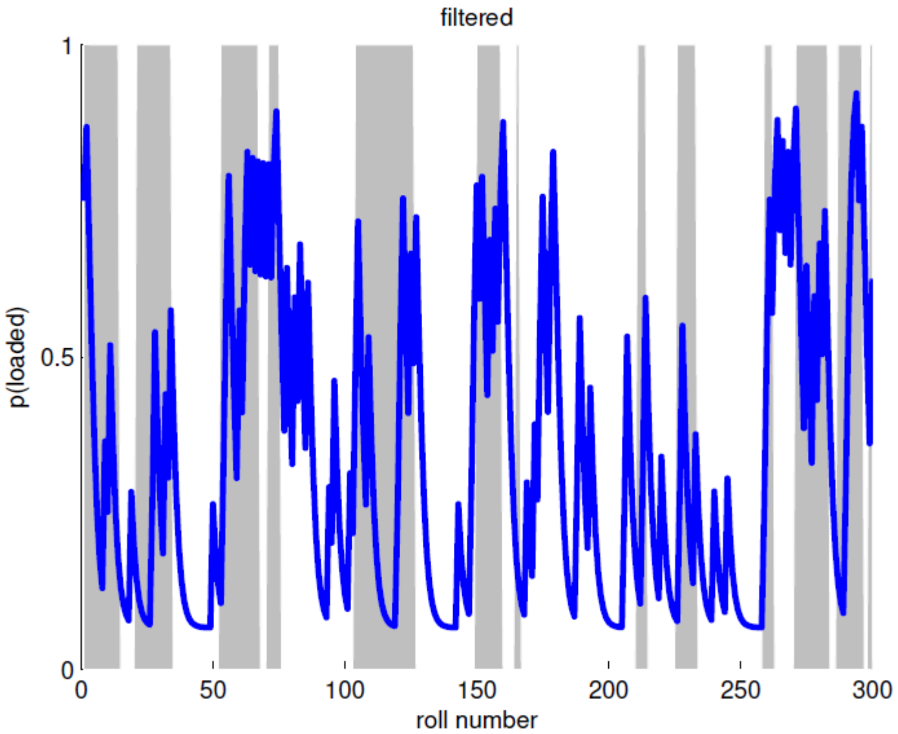
Most probable sequence (decoding)

$$\arg\max_{Y_{1:T}} P(Y_{1:T} \mid X_{1:T}, \theta)$$

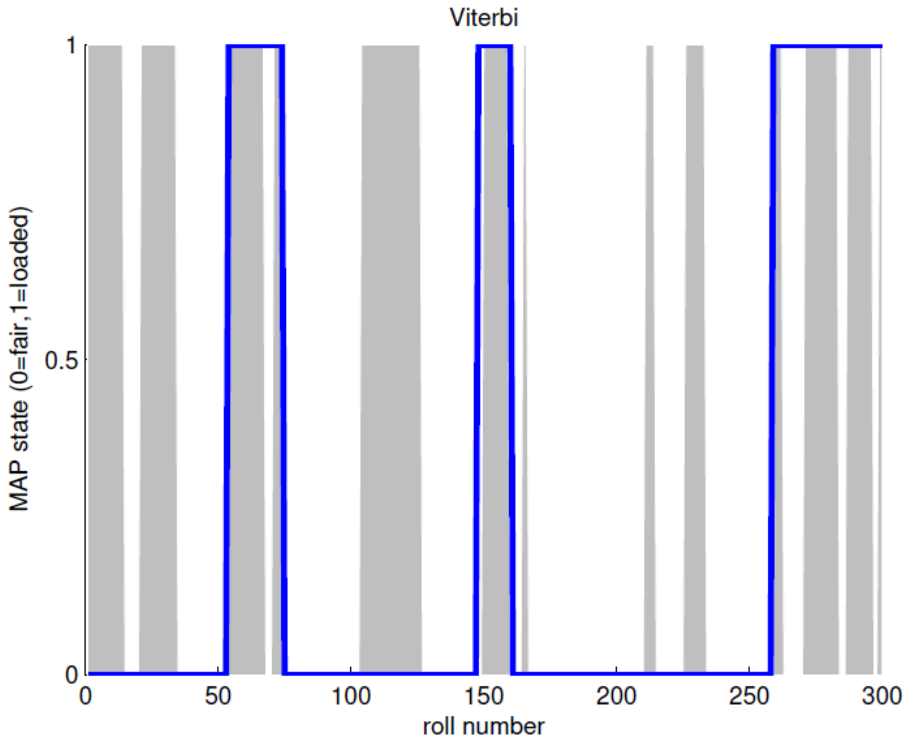
Parameter Estimation

- Given: The observations $X_{1:T}$
- Goal: Find the best parameter estimate of θ

HHM: Example – Dishonest Casino



Conditional probability



Decoding

HHM: Computing the Posterior Probabilities

Filtering Problem

Can directly compute $P(Y_{1:T} \mid X_{1:T}, \theta)$ using the standard way, but that is slow and doesn't exploit the conditional independency structure of HMMs

A popular fast algorithm:

Forward-Backward algorithm, can be done in two passes (one forward pass, one backward pass) over the states.

Decoding Problem

Most likely posterior setting of the hidden states can be computed efficiently using a dynamic programming algorithm, called **Viterbi decoding algorithm**

See supplementary material for detail on these algorithms

HHM: Learning the Parameters

We can use the Expectation Maximization (EM) Algorithm!

Input: *n* observations sequences $x_{1:T}^{(1)},\ x_{1:T}^{(2)},\ \ldots,\ x_{1:T}^{(n)}$

Initialize:

Start with an initial setting / guess of parameters $(\hat{\pi}, \hat{A}, \hat{B})$

E-step:

Compute conditional expectation Y given X and current parameter guess

(this can be done using the Forward-Backward algorithm)

M-step:

Given the estimate of Y and the observations X, we have the complete likelihood, so simply maximize the likelihood by taking the derivative and examine the stationary points.

See supplementary material for details

What We Learned…

- Graphical Models
  - Bayesian Networks and Markov Random Fields
- Doing inference and learning on graphical models
- Markov Models
- Hidden Markov Models
- Bayesian Networks

Questions?