CS221 Problem Workout

Week 7

Content from some slides are inspired by COMS 4701, by Prof. Tony B. Dear of Columbia University

Stanford University

Markov Networks

Markov networks = factor graphs + probability

Definition: Markov network

A Markov network is a factor graph which defines a joint distribution over random variables $X = (X_1, \dots, X_n)$:

$$\mathbb{P}(X = x) = \frac{\text{Weight}(x)}{Z}$$

where $Z = \sum_{x'} \text{Weight}(x')$ is the normalization constant.

CSPs	Markov networks
variables	random variables
weights	probabilities
maximum weight assignment	marginal probabilities

Marginalization

- Given a **joint distribution**, we can find distributions over subsets of
- RVs We can sum out or **marginalize** irrelevant RVs

$$\boxed{P(Y) = \sum_z P(Y, Z = z)}$$

T	W	Pr(T,W)
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$\xrightarrow{P(t) = \sum_w P(t,w)}$$

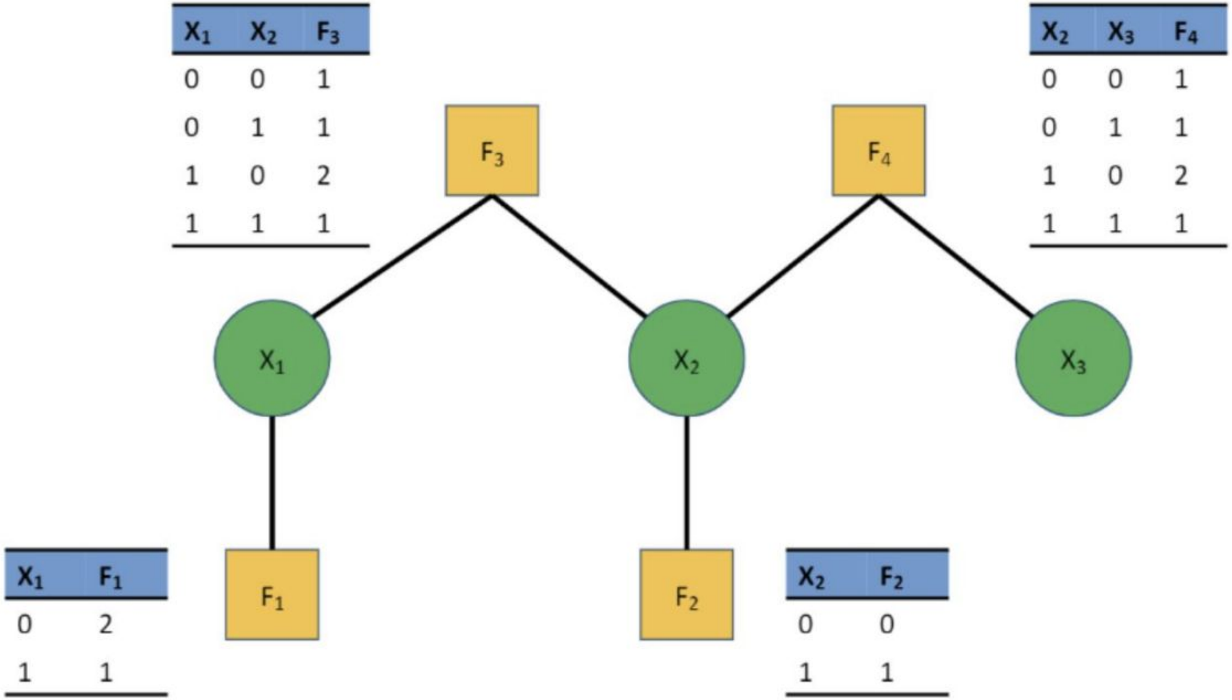
T	Pr(T)
hot	0.5
cold	0.5

$$\xrightarrow{P(w) = \sum_t P(t,w)}$$

W	Pr(W)
sun	0.6
rain	0.4

Problem 1

This problem will give you some practice on computing probabilities given a Markov network. Specifically, given the Markov network below, we will ask you questions about the probability distribution $p(X_1, X_2, X_3)$ over the binary random variables X_1, X_2, and X_3.



X_1	X_2	F_3
0	0	1
0	1	1
1	0	2
1	1	1

X_2	X_3	F_4
0	0	1
0	1	1
1	0	2
1	1	1

X_1	F_1
0	2
1	1

X_2	F_2
0	0
1	1

Car Insurance Pricing

Let’s imagine you are buying car insurance. How does the insurance company come up with a quote given your profile?

Car Insurance Pricing

Let’s imagine you are buying car insurance. How does the insurance company come up with a quote given your profile?

Considerations:

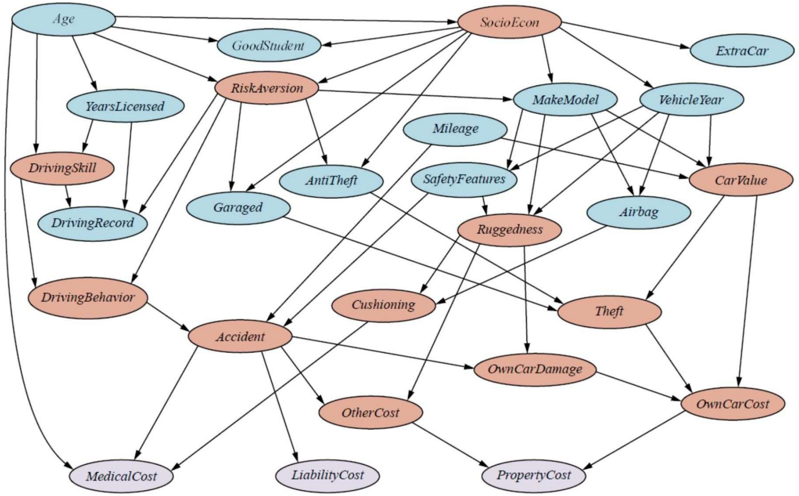
- Pricing model should reflect your driving history, vehicle condition, etc
- Observable variables: age, driving record, vehicle make model.
- Unobservable variables: liability cost, medical cost, etc

Car Insurance Pricing

Let's imagine you are buying car insurance. How does the insurance company come up with a quote given your profile?

Considerations:

- Pricing model should reflect your driving history, vehicle condition, etc
- Observable variables: age, driving record, vehicle make model.
- Unobservable variables: liability cost, medical cost, etc



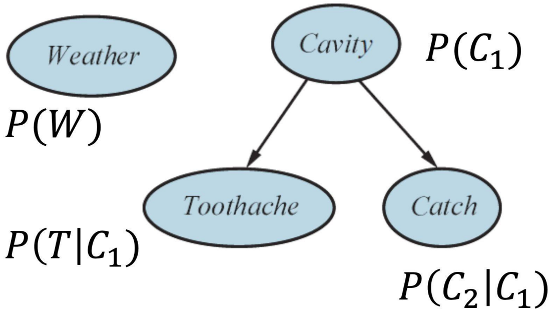
Bayesian Networks

- Handle heterogenously missing information, both at training and test time
- Incorporate prior knowledge (e.g., Mendelian inheritance, laws of physics)
- Can interpret all the intermediate variables
- Precursor to causal models (can do interventions and counterfactuals)

Bayesian Networks

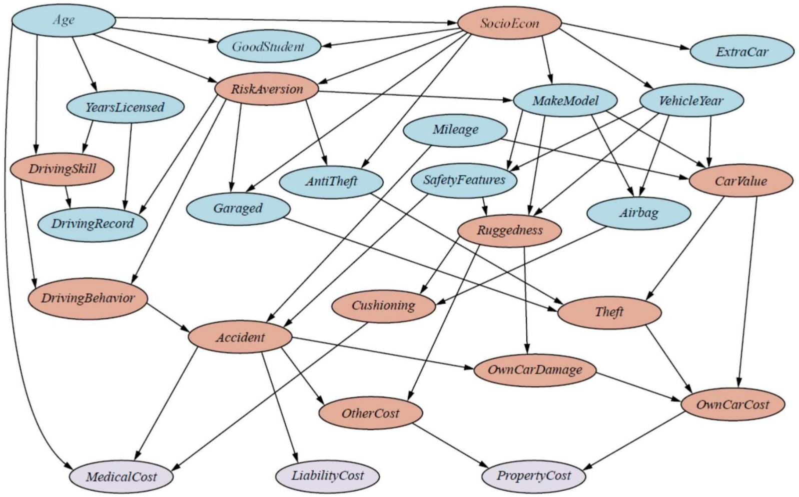
Bayesian network: A directed acyclic graph (DAG) representation of a distribution

- Each node corresponds to a random variable
- Each edge indicates influence or correlation (sometimes causation)
- **Parameters of the Bayes net:** A conditional probability table for each node
- The table for a node *X_i* contains the values *P*(*X_i* | *parents*(*X_i*))



Car Insurance Pricing - Inference

How to compute the **conditional probability** of the **unobservable variables**: liability cost, medical cost, etc, **conditioned on observable variables**: age, driving record, vehicle make model



Bayesian Networks

Joint distribution: we use conditional independence to compute joint distributions.

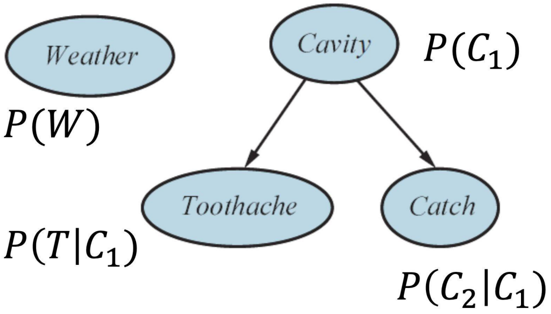
$$P(x_1, \dots, x_n) = \prod_{i=1}^{n} P(x_i | x_1, \dots, x_{i-1}) = \prod_{i=1}^{n} P(x_i | parents(X_i))$$

Bayesian Networks Inference

Joint distribution: we use conditional independence to compute joint distributions.

$$P(x_1, \dots, x_n) = \prod_{i=1}^{n} P(x_i | x_1, \dots, x_{i-1}) = \prod_{i=1}^{n} P(x_i | parents(X_i))$$

- Example:
- $P(w, c1, t, c2) = P(w)\ P(c1)\ P(t \mid c1) P(c2 \mid c1)$

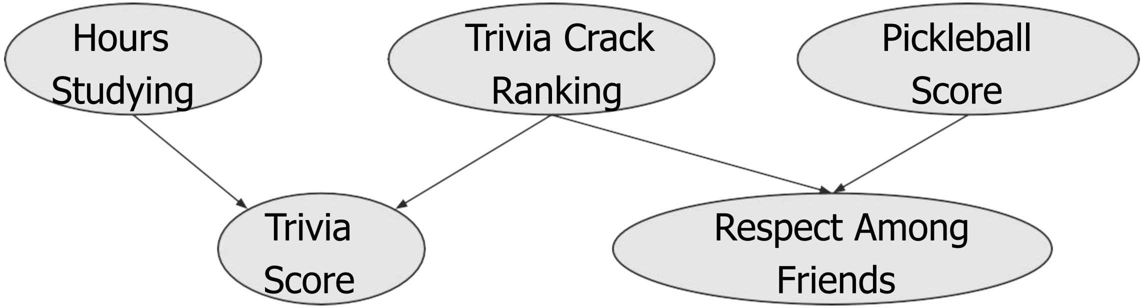


Bayesian Networks Inference

Joint distribution: we use conditional independence to compute joint distributions.

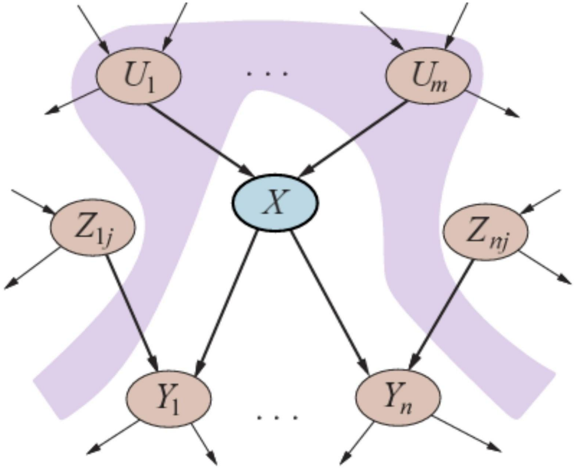
$$P(x_1, \dots, x_n) = \prod_{i=1}^{n} P(x_i | x_1, \dots, x_{i-1}) = \prod_{i=1}^{n} P(x_i | parents(X_i))$$

- Structure of the Bayes Net reveals relations between variables.
- If we are given a table of P(Trivia Score | Hours Studying), it is likely the case that **Hour Studying is parent of trivia score**!



Conditional Independence

- We know that a node is independent of its "ancestors" given all its parents
- More generally, **a node is independent of its "non-descendants" given its parents**
- These imply several local conditional independences that can be inferred from Bayes net structure only



Probability essentials

- Conditional probability $P(x|y) = \dfrac{P(x,y)}{P(y)}$
- Product rule $P(x,y) = P(x|y)P(y)$
- Chain rule
$$\begin{aligned} P(X_1, X_2, \dots X_n) &= P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)\,.. \\ &= \prod_{i=1}^{n} P(X_i|X_1, \dots, X_{i-1}) \end{aligned}$$
- X, Y are independent iff: $\forall x, y : P(x,y) = P(x)P(y)$
- X and Y are conditionally independent given Z iff:

$$\forall x, y, z : P(x, y|z) = P(x|z)P(y|z) \qquad X \perp\!\!\!\perp Y | Z$$

- Bayes rule $P(x|y) = \dfrac{P(x,y)}{P(y)} = \dfrac{P(y|x)P(x)}{P(y)}$

Conditional probability example

- Someone's footsize is correlated to their literacy: older people have bigger feet and are more likely to be literate.
- Once you condition on age though, this relationship disappears.
- Footsize and literacy are conditionally independent given an age.

Source: https://math.stackexchange.com/a/3854506

https://www.offthewagonshop.com/products/office-bigfoot

Problem 2

$P(A\|D,X)$			
$+d$	$+x$	$+a$	0.9
$+d$	$+x$	$-a$	0.1
$+d$	$-x$	$+a$	0.8
$+d$	$-x$	$-a$	0.2
$-d$	$+x$	$+a$	0.6
$-d$	$+x$	$-a$	0.4
$-d$	$-x$	$+a$	0.1
$-d$	$-x$	$-a$	0.9

$P(D)$	
$+d$	0.1
$-d$	0.9

$P(X\|D)$		
$+d$	$+x$	0.7
$+d$	$-x$	0.3
$-d$	$+x$	0.8
$-d$	$-x$	0.2

$P(B\|D)$		
$+d$	$+b$	0.7
$+d$	$-b$	0.3
$-d$	$+b$	0.5
$-d$	$-b$	0.5

(a) Given the tables above, draw a minimal representative Bayesian network of this model. Be sure to label all nodes and the directionality of the edges.

(b) Compute the following probabilities: $P(+d, +a)$, $P(+d \mid +a)$, $P(+d \mid +b)$.

(c) Which of the following conditional independences are guaranteed by the above network?

□ $X \perp\!\!\!\perp B \mid D$

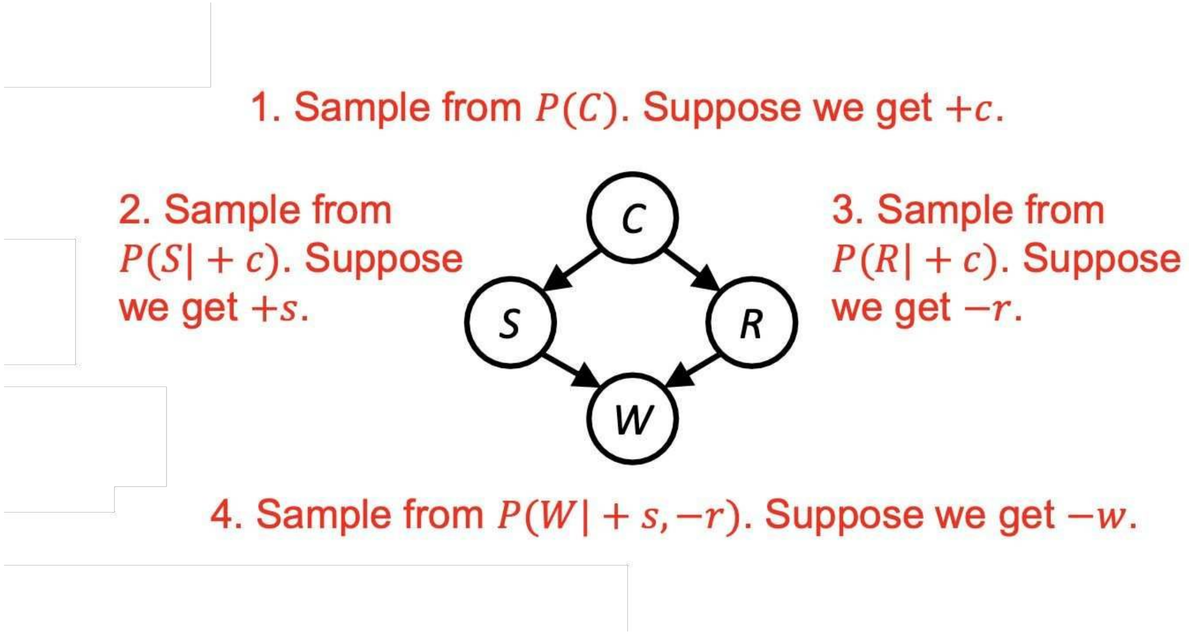
□ $D \perp\!\!\!\perp A \mid B$

□ $D \perp\!\!\!\perp A \mid X$

□ $D \perp\!\!\!\perp X \mid A$

Sampling

- **Motivation:** Exact inference becomes impossible when we have too many variables
- **Sample** the Bayes net using the known conditional probability tables



1. Sample from $P(C)$. Suppose we get $+c$.

2. Sample from $P(S|+c)$. Suppose we get $+s$.

3. Sample from $P(R|+c)$. Suppose we get $-r$.

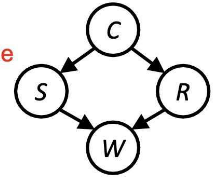
4. Sample from $P(W|+s,-r)$. Suppose we get $-w$.

Sampling

- **Motivation**: Exact inference becomes impossible when we have too many variables
- **Sample** the Bayes net using the known conditional probability tables

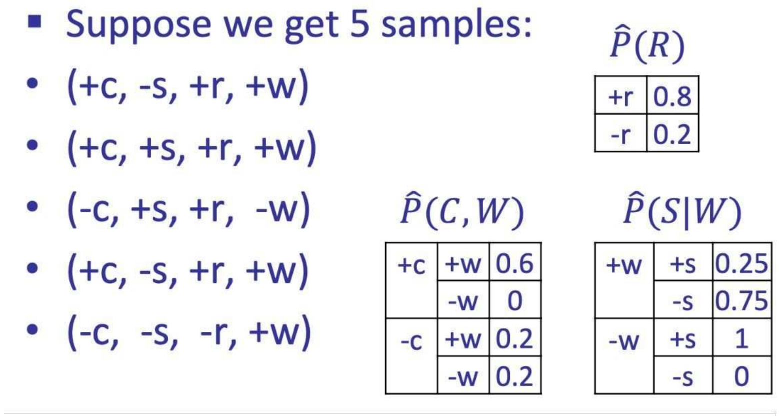
1. Sample from $P(C)$. Suppose we get $+c$.

2. Sample from $P(S|+c)$. Suppose we get $+s$.



3. Sample from $P(R|+c)$. Suppose we get $-r$.

4. Sample from $P(W|+s,-r)$. Suppose we get $-w$.



- Suppose we get 5 samples:
- (+c, -s, +r, +w)
- (+c, +s, +r, +w)
- (-c, +s, +r, -w)
- (+c, -s, +r, +w)
- (-c, -s, -r, +w)

$\hat{P}(R)$

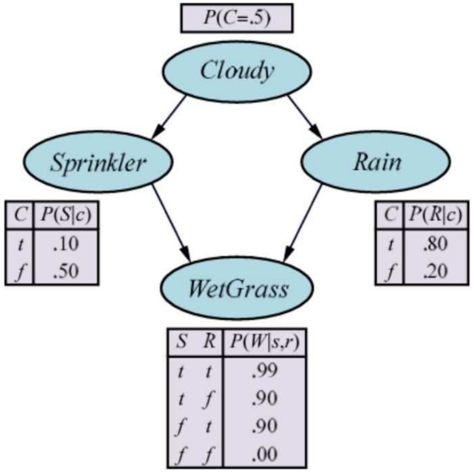
+r	0.8
-r	0.2

$\hat{P}(C,W)$

+c	+w	0.6
	-w	0
-c	+w	0.2
	-w	0.2

$\hat{P}(S|W)$

+w	+s	0.25
	-s	0.75
-w	+s	1
	-s	0



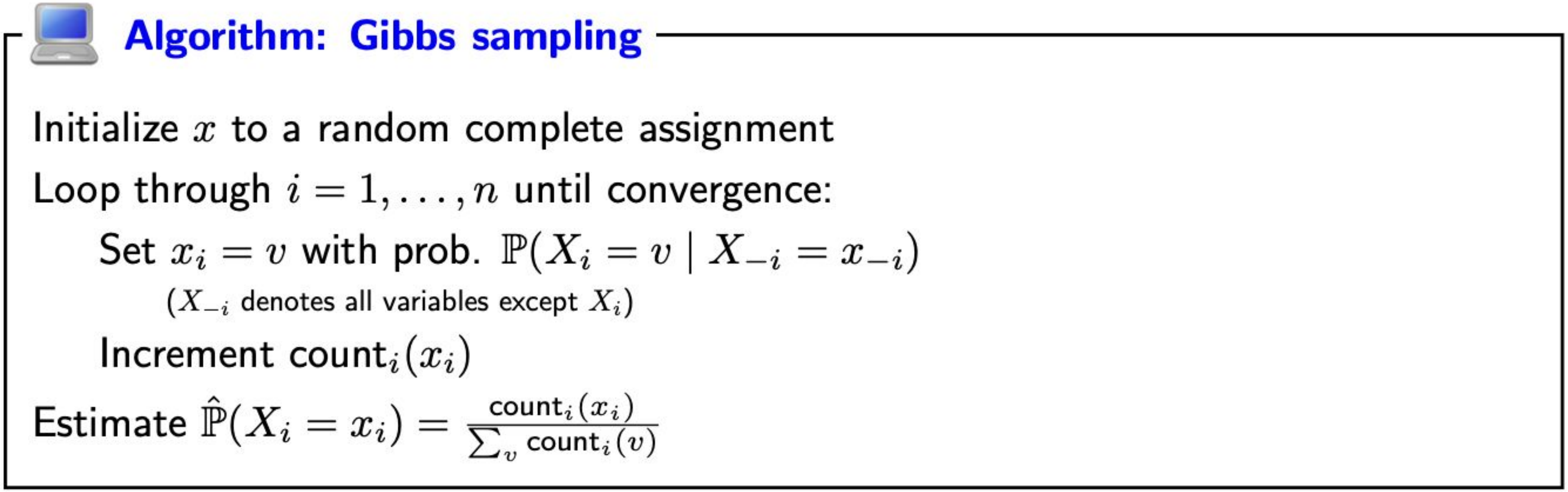
C	P(S|c)
t	.10
f	.50

C	P(R|c)
t	.80
f	.20

S	R	P(W|s,r)
t	t	.99
t	f	.90
f	t	.90
f	f	.00

Gibbs Sampling

- **Problem:** How do we sample from $P(Xi$ | all other nodes in Bayes Net)?



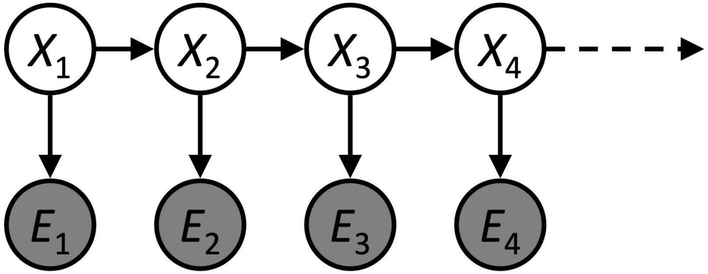
Algorithm: Gibbs sampling

```
Initialize $x$ to a random complete assignment
Loop through $i = 1, \ldots, n$ until convergence:
    Set $x_i = v$ with prob. $\mathbb{P}(X_i = v \mid X_{-i} = x_{-i})$
        ($X_{-i}$ denotes all variables except $X_i$)
    Increment $\text{count}_i(x_i)$
Estimate $\hat{\mathbb{P}}(X_i = x_i) = \frac{\text{count}_i(x_i)}{\sum_v \text{count}_i(v)}$
```

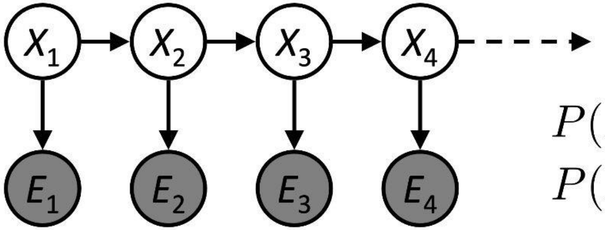
Special Case of Bayes Net: HMM

- **Hidden Markov model:** A Markov process with hidden states X_t and observable evidence variables E_t

- Transition model: $P(X\_t|X\_t-1)$
- Observation model: $P(E\_t|X\_t)$



HMM Inference



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

- General joint distribution:

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

- Marginal distributions can be found by summing out RVs
- For certain computations we don’t even need the entire joint distribution!

Problem 3

The viewerships of the two teams evolve according to the following model, where each month a fan is either gained or lost with equal probability:

$$\Pr(M_{t+1}|M_t) = \begin{cases} \frac{1}{2} & \text{if } M_{t+1} = M_t - 1 \\ \frac{1}{2} & \text{if } M_{t+1} = M_t + 1 \\ 0 & \text{otherwise} \end{cases} \qquad \Pr(B_{t+1}|B_t) = \begin{cases} \frac{1}{2} & \text{if } B_{t+1} = B_t - 1 \\ \frac{1}{2} & \text{if } B_{t+1} = B_t + 1 \\ 0 & \text{otherwise} \end{cases}$$

The Bayesian fans like to rewatch their trivia shows by searching the recaps online! We model the fan's size's influence on the number of internet searches by:

$$\Pr(S_t|B_t) = \begin{cases} 0.3 & \text{if } S_t = B_t \\ 0.25 & \text{if } S_t = B_t - 1 \\ 0.2 & \text{if } S_t = B_t - 2 \\ 0.15 & \text{if } S_t = B_t - 3 \\ 0.1 & \text{if } S_t = B_t - 4 \\ 0 & \text{otherwise} \end{cases}$$

Lastly, because most TV viewers attend each monthly friendly matches (although sometimes more, and sometimes fewer), we model the influence of the TV viewership number on the friendly match attendance by:

$$\Pr(A_t|B_t, M_t) = \begin{cases} 0.14 & \text{if } A_t = B_t + M_t \\ 0.13 & \text{if } |A_t - (B_t + M_t)| = 1 \\ 0.11 & \text{if } |A_t - (B_t + M_t)| = 2 \\ 0.09 & \text{if } |A_t - (B_t + M_t)| = 3 \\ 0.06 & \text{if } |A_t - (B_t + M_t)| = 4 \\ 0.04 & \text{if } |A_t - (B_t + M_t)| = 5 \\ 0 & \text{otherwise} \end{cases}$$

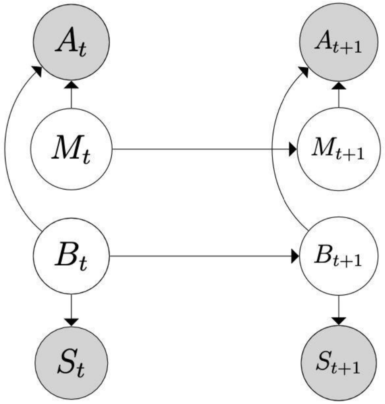


Figure 1: The changing TV viewership count modeled as a dynamic Bayesian network. The unshaded nodes correspond to the latent/hidden TV viewership counts, and the shaded nodes correspond to the observable emissions.

a. (*10 points*) **Inference**

Suppose the Bayesian's trivia team captain took a nationwide poll in month t that concluded they had exactly 75 TV viewers. Suppose additionally that in month $t+2$, the search engine reported 73 people search for the Bayesians online. What is the probability that in month $t+2$ the Bayesians have 77 TV viewers?

$\Pr(B_{t+2} = 77 | B_t = 75, S_{t+2} = 73) =$

Problem 3

The viewerships of the two teams evolve according to the following model, where each month a fan is either gained or lost with equal probability:

$$\Pr(M_{t+1}|M_t) = \begin{cases} \frac{1}{2} & \text{if } M_{t+1} = M_t - 1 \\ \frac{1}{2} & \text{if } M_{t+1} = M_t + 1 \\ 0 & \text{otherwise} \end{cases} \qquad \Pr(B_{t+1}|B_t) = \begin{cases} \frac{1}{2} & \text{if } B_{t+1} = B_t - 1 \\ \frac{1}{2} & \text{if } B_{t+1} = B_t + 1 \\ 0 & \text{otherwise} \end{cases}$$

The Bayesian fans like to rewatch their trivia shows by searching the recaps online! We model the fan's size's influence on the number of internet searches by:

$$\Pr(S_t|B_t) = \begin{cases} 0.3 & \text{if } S_t = B_t \\ 0.25 & \text{if } S_t = B_t - 1 \\ 0.2 & \text{if } S_t = B_t - 2 \\ 0.15 & \text{if } S_t = B_t - 3 \\ 0.1 & \text{if } S_t = B_t - 4 \\ 0 & \text{otherwise} \end{cases}$$

Lastly, because most TV viewers attend each monthly friendly matches (although sometimes more, and sometimes fewer), we model the influence of the TV viewership number on the friendly match attendance by:

$$\Pr(A_t|B_t, M_t) = \begin{cases} 0.14 & \text{if } A_t = B_t + M_t \\ 0.13 & \text{if } |A_t - (B_t + M_t)| = 1 \\ 0.11 & \text{if } |A_t - (B_t + M_t)| = 2 \\ 0.09 & \text{if } |A_t - (B_t + M_t)| = 3 \\ 0.06 & \text{if } |A_t - (B_t + M_t)| = 4 \\ 0.04 & \text{if } |A_t - (B_t + M_t)| = 5 \\ 0 & \text{otherwise} \end{cases}$$

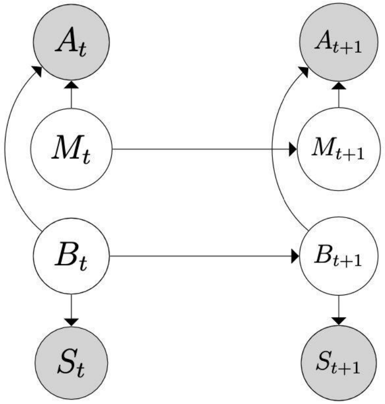


Figure 1: The changing TV viewership count modeled as a dynamic Bayesian network. The unshaded nodes correspond to the latent/hidden TV viewership counts, and the shaded nodes correspond to the observable emissions.

b. (*4 points*) **Extra Practice - Gibbs Sampling**

Inference is exhausting; you decide that you'd be satisfied with simply being able to draw samples from distributions rather than specifying them exactly. In particular, you want to sample joint assignments to the variables $\{B_t, M_t, A_t, S_t\}_{t=1}^{T}$ for some time horizon T. You decide to implement Gibbs sampling for this purpose, but something's not right! What additional information, beyond what we've given you, would allow you to perform Gibbs sampling? Briefly explain.

Thank You