

Machine Learning

4771

Instructor: Tony Jebara

Topic 15

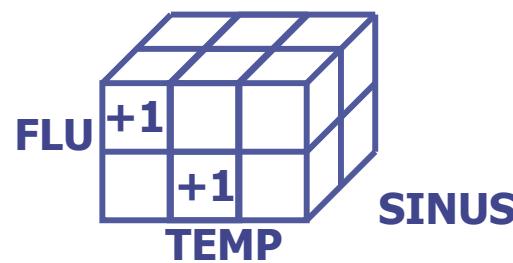
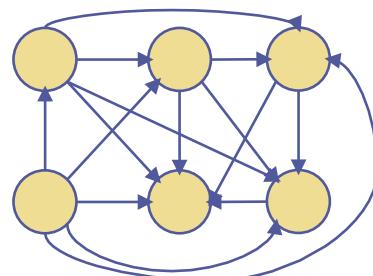
- Graphical Models
- Maximum Likelihood for Graphical Models
- Testing for Conditional Independence & D-Separation
- Bayes Ball

Learning Fully Observed Models

- Easiest scenario: we have observed all the nodes
- Want to learn the probability tables from data...
- Have N iid patients:

| PATIENT | FLU | FEVER | SINUS | TEMP | SWELL | HEAD |
|---------|-----|-------|-------|------|-------|------|
| 1 | Y | Y | N | L | Y | Y |
| 2 | N | N | N | M | N | Y |
| 3 | Y | N | Y | H | Y | N |
| 4 | Y | N | Y | M | N | N |

- Simplest case: least general graph handle each dim individually as Bernoulli/Multinomial
- 2nd Simplest case: most general, count each entry in pdf



Divide by total count
Since $\sum_{x_1} \dots \sum_{x_6} p(x) = 1$

- What about learning graphs in between?

Maximum Likelihood CPTs

- Each conditional probability table θ_i part of our parameters
- Given table, have pdf

$$p(X_U | \theta) = \prod_{i=1}^M p(x_i | \pi_i, \theta_i)$$

- Have M variables:

$$X_U = \{x_1, \dots, x_M\}$$
 M variables

- Have N x M dataset:

$$\mathcal{D} = \{X_{U,1}, \dots, X_{U,N}\}$$
 N samples

- Maximum likelihood:

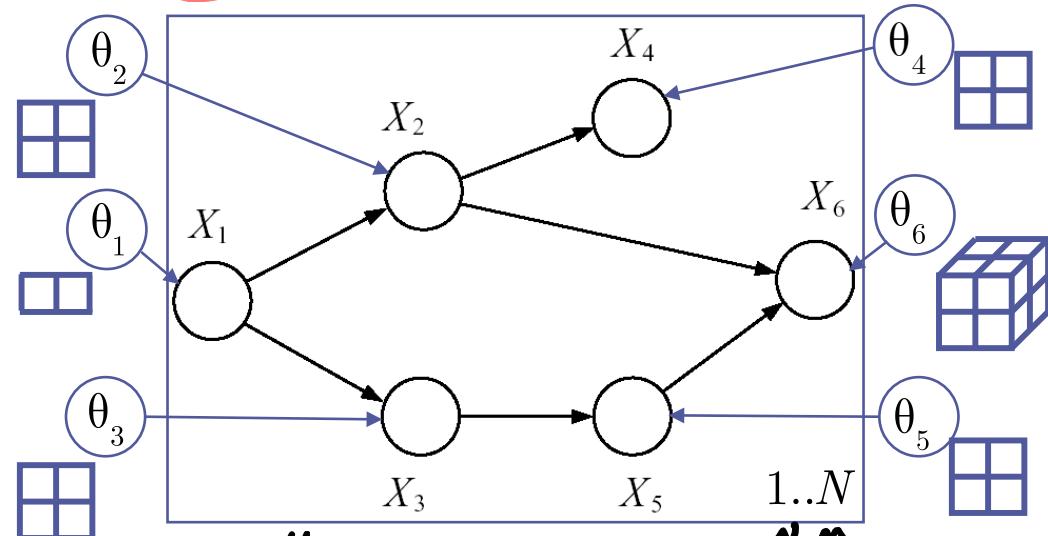
$$\theta^* = \arg \max_{\theta} \log p(\mathcal{D} | \theta) = \arg \max_{\theta} \log \prod_{i=1}^N p(X_{U,i} | \theta) = \arg \max_{\theta} \log \prod_{i=1}^N \prod_{j=1}^M p(x_{i,j} | \pi_{i,j}, \theta_i)$$

\mathcal{D} - Dataset

$$= \arg \max_{\theta} \sum_{n=1}^N \log p(X_{U,n} | \theta)$$

$$= \arg \max_{\theta} \sum_{n=1}^N \log \prod_{i=1}^M p(x_{i,n} | \pi_{i,n}, \theta_i)$$

$$= \arg \max_{\theta} \sum_{n=1}^N \sum_{i=1}^M \log p(x_{i,n} | \pi_{i,n}, \theta_i)$$



each θ_i appears independently, can do ML for each CPT alone! efficient storage & efficient learning

Maximum Likelihood CPTs

Sample n, Sample m

$$\delta(X_{U,n}, X_{U,m}) = \begin{cases} 1 & \text{if } X_{U,n} = X_{U,m} \\ 0 & \text{otherwise} \end{cases}$$

Variables Π are the same: patient n and m have the same symptoms Π
 \uparrow
 Π is the set of symptoms

Counts: # of times what's in the bracket appeared in data, for example:

$$N = \sum_{x_1} m(x_1) = \sum_{x_1} \left(\sum_{x_2} m(x_1, x_2) \right) = \sum_{x_1} \left(\sum_{x_2} \left(\sum_{x_3} m(x_1, x_2, x_3) \right) \right)$$

• So... $l(\theta) = \sum_{n=1}^N \log p(X_{U,n} | \theta)$

variable X_U

$$\sum_{n=1}^N \delta(X_{U,n}, X_{U,n}) = m(X_U)$$

$$= \sum_{n=1}^N \log \prod_{X_U} p(X_U | \theta)^{\delta(X_U, X_{U,n})}$$

$$= \sum_{n=1}^N \sum_{X_U} \delta(X_U, X_{U,n}) \log p(X_U | \theta)$$

$$= \sum_{X_U} m(X_U) \log p(X_U | \theta) = \sum_{X_U} m(X_U) \log \prod_{i=1}^M p(x_i | \pi_i, \theta_i)$$

$$= \sum_{X_U} \sum_{i=1}^M m(X_U) \log p(x_i | \pi_i, \theta_i)$$

number of samples have the same variable x_i (The number of patients have the same symptom x_i)

$$\begin{aligned} m(x_i) &= \sum_{n=1}^N \delta(x_i, x_{i,n}) \\ m(X_U) &= \sum_{n=1}^N \delta(X_U, X_{U,n}) \\ m(X_C) &= \sum_{X_{U \setminus C}} m(X_U) \end{aligned}$$

The number of patients have a subset of symptoms the same as X_C (X_C : a set of variables)

$$\sum_{X_U \setminus X_i \setminus \pi_i} m(X_U) = m(X_i, \pi_i)$$

Tony Jebara, Columbia University

Maximum Likelihood CPTs

- Continuing:
$$l(\theta) = \sum_{X_U} \sum_{i=1}^M m(X_U) \log p(x_i | \pi_i, \theta_i)$$

$$= \sum_{i=1}^M \sum_{x_i, \pi_i} \sum_{X_{U \setminus x_i \setminus \pi_i}} m(X_U) \log p(x_i | \pi_i, \theta_i)$$

$$\pi_i = \text{pa}_i = \text{parent of } i = \sum_{i=1}^M \sum_{x_i, \pi_i} \overline{m(x_i, \pi_i)} \log p(x_i | \pi_i, \theta_i)$$

- Define: $\underline{\theta(x_i, \pi_i) = p(x_i | \pi_i, \theta_i)}$ Constraint: $\sum_{x_i} \theta(x_i, \pi_i) = 1$

- Now have above with Lagrange multipliers:

$$l(\theta) = \sum_{i=1}^M \sum_{x_i} \sum_{\pi_i} m(x_i, \pi_i) \log \theta(x_i, \pi_i) - \sum_{i=1}^M \sum_{\pi_i} \lambda_{\pi_i} \left(\sum_{x_i} \theta(x_i, \pi_i) - 1 \right)$$

$$\frac{\partial l(\theta)}{\partial \theta(x_i, \pi_i)} = \frac{\cancel{m(x_i, \pi_i)}}{\theta(x_i, \pi_i)} - \lambda_{\pi_i} = 0 \rightarrow \theta(x_i, \pi_i) = \frac{m(x_i, \pi_i)}{\lambda_{\pi_i}}$$

- Plug constraint: $\sum_{x_i} \frac{m(x_i, \pi_i)}{\lambda_{\pi_i}} = 1 \rightarrow \lambda_{\pi_i} = \sum_{x_i} m(x_i, \pi_i) = m(\pi_i)$

- Final solution (trivial!):

$$\theta(x_i, \pi_i) = \frac{m(x_i, \pi_i)}{m(\pi_i)}$$

Maximum Likelihood CPTs

- Continuing:
$$\begin{aligned} l(\theta) &= \sum_{X_U} \sum_{i=1}^M m(X_U) \log p(x_i | \pi_i, \theta_i) \\ &= \sum_{i=1}^M \sum_{x_i, \pi_i} \sum_{X_{U \setminus x_i \setminus \pi_i}} m(X_U) \log p(x_i | \pi_i, \theta_i) \\ &= \sum_{i=1}^M \sum_{x_i, \pi_i} \overbrace{m(x_i, \pi_i)}^{\text{log}} \log p(x_i | \pi_i, \theta_i) \end{aligned}$$

- Define: $\theta(x_i, \pi_i) = p(x_i | \pi_i, \theta_i)$ Constraint: $\sum_{x_i} \theta(x_i, \pi_i) = 1$

- Now have above with Lagrange multipliers:

$$\begin{aligned} l(\theta) &= \sum_{i=1}^M \sum_{x_i} \sum_{\pi_i} m(x_i, \pi_i) \log \theta(x_i, \pi_i) - \sum_{i=1}^M \sum_{\pi_i} \lambda_{\pi_i} \left(\sum_{x_i} \theta(x_i, \pi_i) - 1 \right) \\ \frac{\partial l(\theta)}{\partial \theta(x_i, \pi_i)} &= \frac{m(x_i, \pi_i)}{\theta(x_i, \pi_i)} - \lambda_{\pi_i} = 0 \rightarrow \theta(x_i, \pi_i) = \frac{m(x_i, \pi_i)}{\lambda_{\pi_i}} \end{aligned}$$

- Plug constraint: $\sum_{x_i} \frac{m(x_i, \pi_i)}{\lambda_{\pi_i}} = 1 \rightarrow \lambda_{\pi_i} = \sum_{x_i} m(x_i, \pi_i) = m(\pi_i)$

- Final solution (trivial!):

$$\theta(x_i, \pi_i) = \frac{m(x_i, \pi_i) + \varepsilon}{m(\pi_i) + \varepsilon |x_i|}$$

**MAP
VERSION**

Maximum Likelihood CPTs

- Let's try an example:
- Compute the cpt

$$p(x_3 | x_1)$$

| PATIENT | x_1 | x_3 | FLU | FEVER | SINUS | TEMP | SWELL | HEAD |
|---------|-------|-------|-----|-------|-------|------|-------|------|
| 1 | Y | Y | N | L | Y | Y | | |
| 2 | N | N | N | M | N | Y | | |
| 3 | Y | N | Y | H | Y | N | | |
| 4 | Y | N | Y | M | N | N | | |

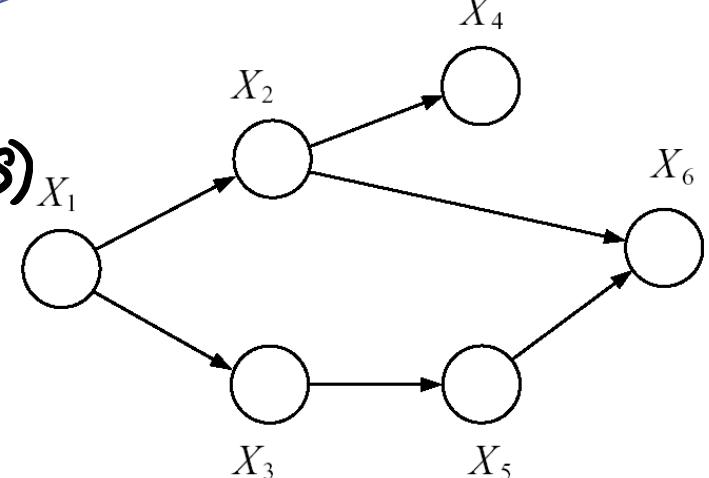
- Using the formula:

$$\theta(x_i, \pi_i) = \frac{m(x_i, \pi_i)}{m(\pi_i)}$$

Note, here 0/0 = prior constant

| | | |
|-----------|-----------|-----------------|
| $x_1 = 0$ | $x_1 = 1$ | $m(FLU, SINUS)$ |
| $x_3 = 0$ | 1 | |
| 0 | 2 | $m(x_3, x_1)$ |
| $x_3 = 1$ | 1 | $m(x_1)$ |

$$\begin{aligned} \frac{1}{1} &= \frac{1}{1+2} = \frac{1}{3} \\ \frac{0}{1} &= \frac{2}{1+2} = \frac{2}{3} \end{aligned} \quad p(x_3 | x_1)$$



Efficient, only count over
subset of variables in $p(X_B | X_A)$
Not all $p(x_1, \dots, x_M)$

Conditional Dependence Tests

- Another thing we would like to do with a graphical model:
Check conditional independencies...

"Is Temperature Independ. of Flu Given Fever?"

"Is Temperature Independ. of Sinus Infection Given Fever?"

- Try computing & simplify marginals of $p(x)$

$$p(X) = p(x_1)p(x_2 | x_1)p(x_3 | x_1)p(x_4 | x_2)p(x_5 | x_3)p(x_6 | x_2, x_5)$$

$$p(x_4 | x_1, x_2, x_3) = \frac{p(x_1, x_2, x_3, x_4)}{p(x_1, x_2, x_3)} = \frac{\sum_{x_5} \sum_{x_6} p(X)}{\sum_{x_4} \sum_{x_5} \sum_{x_6} p(X)}$$

$$= \frac{p(x_1)p(x_2 | x_1)p(x_3 | x_1)p(x_4 | x_2)}{p(x_1)p(x_2 | x_1)p(x_3 | x_1)}$$

$$= p(x_4 | x_2) \quad \xrightarrow{x_4 \perp\!\!\!\perp x_1, x_3 | x_2}$$

①
Easy Case.

②
Harder Case

- In this case it was easy, what if checking: $x_1 \perp\!\!\!\perp x_6 | x_2, x_3$
- Hard to compute $p(x_1 | x_2, x_3, x_6)$ want efficient algorithm...

D-Separation & Bayes Ball

- There is a graph algorithm for checking independence
- Intuition: separation or blocking of some nodes by others

incorrect for
directed.
 correct for
undirected.

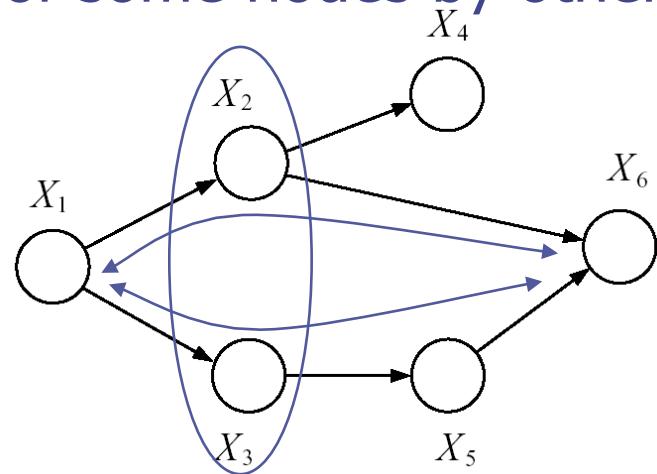
- Example:

if nodes x_2, x_3 “block”
path from x_1 to x_6
we might say that

$$x_1 \perp\!\!\!\perp x_6 \mid x_2, x_3$$

- This is not exact for directed graphs (true for Undirected)
- We need more than just simple Separation
- Need D-Separation (directed separation)
- D-Separation is computed via the Bayes Ball algorithm
- Use to prove general statements over subsets of vars:

$$X_A \perp\!\!\!\perp X_B \mid X_C$$



Bayes Ball Algorithm

- The algorithm:

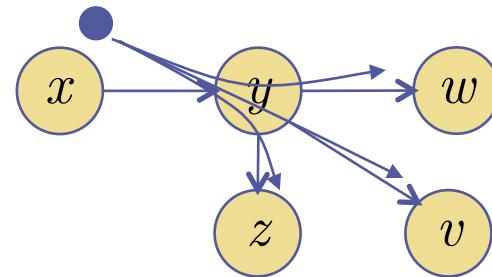
$$X_A \perp\!\!\!\perp X_B \mid X_C$$

- 1) Shade nodes X_C
- 2) Place a ball at each node in X_A
- 3) Bounce balls around graph according to some *rules*
- 4) If no balls reach X_B , then $X_A \perp\!\!\!\perp X_B \mid X_C$ is true (else false)

Balls can travel along/against arrows

Pick any incoming & outgoing path

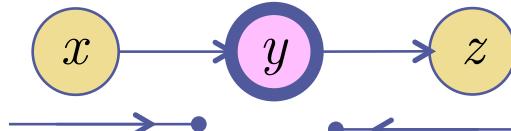
Test each to see if ball goes through or bounces back



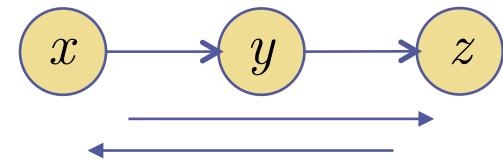
Look at canonical sub-graphs & leaf cases for rules...

Bayes Ball Algorithm

1) Markov Chain:



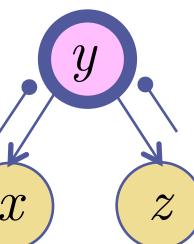
Rule depends only on shading of middle node



Go Through $x \text{ } \cancel{\underline{x}} \text{ } z$

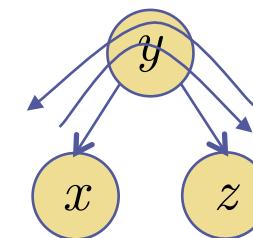
2) Two Effects:

Ball stops $x \parallel z | y$



Rule depends only on shading of middle node

Ball stops $x \parallel z | y$

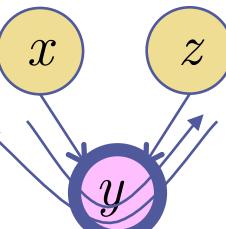


Go Through $x \text{ } \cancel{\underline{x}} \text{ } z$

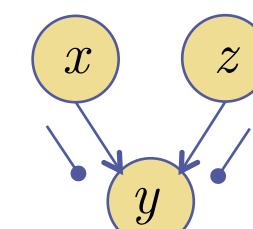
3) Two Causes (V):

~~No Pass~~

Rule depends only on shading of middle node



Go Through $x \text{ } \cancel{\underline{x}} \text{ } z | y$

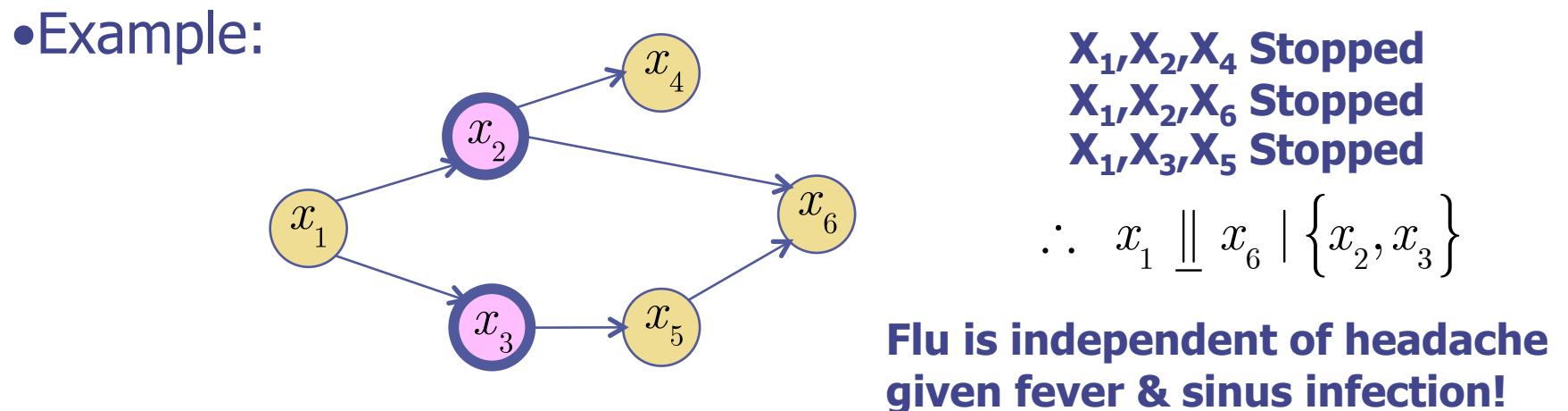
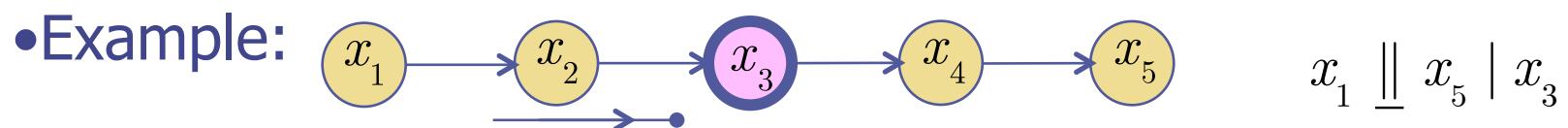
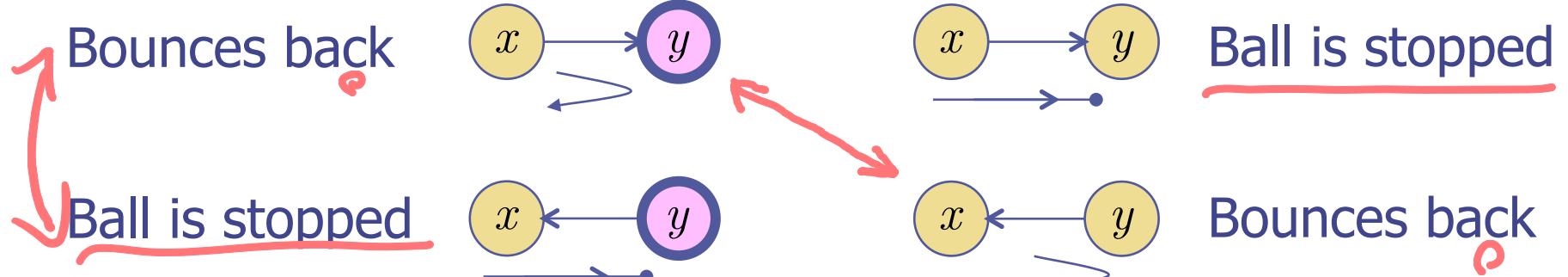


Ball stops $x \parallel z$



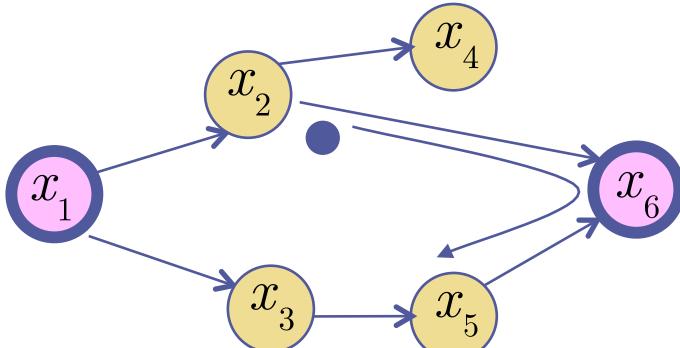
Bayes Ball Algorithm

- Also need to look at special 'leaf' cases:

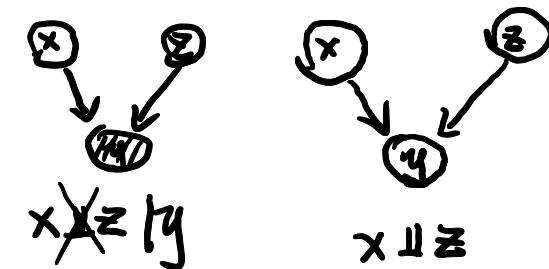
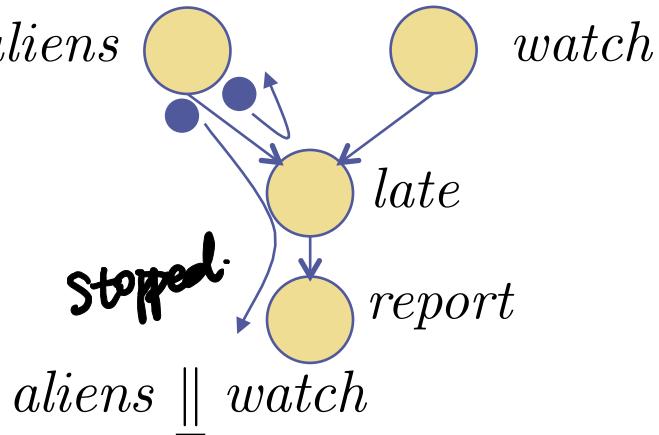


Bayes Ball Algorithm

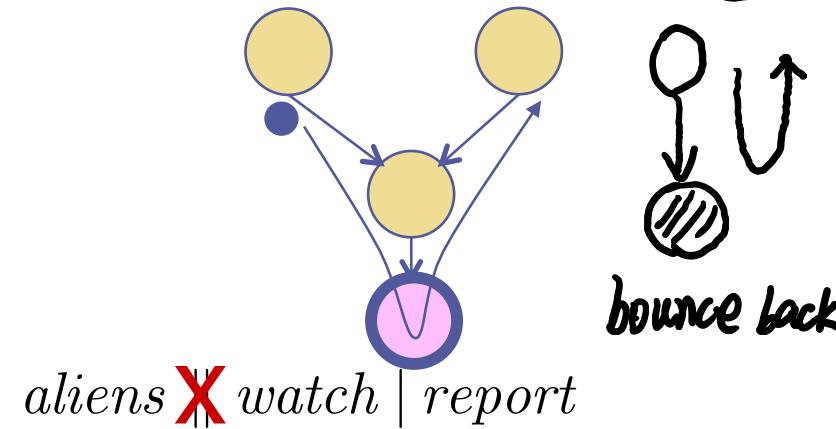
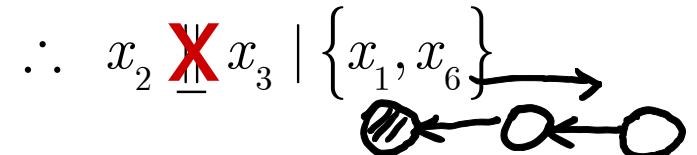
- Example:



- Example: aliens watch



x_2, x_6, x_5 Goes Through
Because of V-structure



Ball bounces back from report leaf and goes to right if report is shaded. Bob is waiting for Alice but can't know if she is late. Instead a security guard says if she is. She can be late if aliens abduct her or Bob's watch is ahead (daylight savings time). Guard reports she is late. If watch is ahead, $p(\text{alien}=\text{true})$ goes down, they are dependent.