



B: 8.4  
KF, Chapter 15 – 15.5

---

# Learning Belief Net Structures

R Greiner


Cmput 466 / 551

Some material taken from C Guesterin (CMU), K Murphy (UBC)



# Outline

---

- Motivation
  - What is a Belief Net?
  - Learning a Belief Net
    - Goal?
    - Learning Parameters – Complete Data
    - Learning Parameters – Incomplete Data
    - Learning Structure
- 

# Learning Belief Nets

Structure

Known

Unknown

Data

Complete

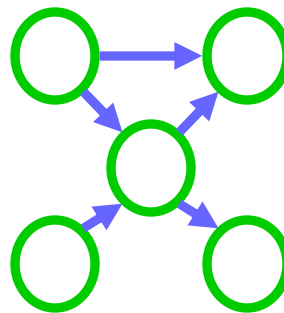
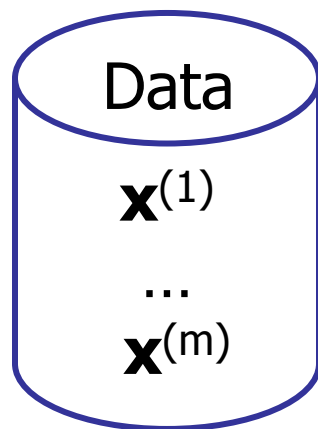
✓ **Easy**

**NP-hard**

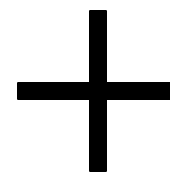
Missing

✓ **Hard ... EM**

**Very hard!!**



**structure**



CPTs :

$P(X_i | \mathbf{Pa}_{X_i})$

**parameters**

# Learning the structure of a BN



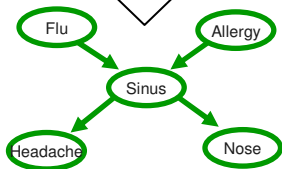
Data

$\langle x_1^{(1)}, \dots, x_n^{(1)} \rangle$

...

$\langle x_1^{(m)}, \dots, x_n^{(m)} \rangle$

Learn structure and parameters



## ■ Constraint-based approach

- BN encodes conditional independencies
- Test conditional independencies in data
- Find an I-map (?P-map?)

## ■ Score-based approach

- Finding structure + parameters is *density estimation*
- Evaluate model as we evaluated parameters
  - Maximum likelihood
  - Bayesian
  - etc.



# Remember: Obtaining a P-map?

---

- Given  $I(P)$  = independence assertions that are true for  $P$ 
  1. Obtain skeleton
  2. Obtain immoralities
  3. Using skeleton and immoralities, obtain every (and any) BN structure from the equivalence class

## ■ **Constraint-based approach:**

- Use Learn\_PDAG algorithm
- Key question: **Independence test**



# Independence tests

---

- Statistically difficult task!
- Intuitive approach: **Mutual information**

$$I(X, Y) = \sum_{x,y} P(x, y) \log \frac{P(x, y)}{P(x)P(y)}$$

- Mutual information and independence:
  - X and Y independent if and only if  $I(X, Y) = 0$
  - $X \perp Y \Rightarrow P(x, y) = P(x)P(y) \Rightarrow \log[ P(x, y)/P(x)P(y) ] = 0$

- **Conditional mutual information:**

$$I(X, Y|Z) = E_Z[ I[X, Y|Z = z] ] = \sum_z \sum_{x,y} P(x, y|z) \log \frac{P(x, y|z)}{P(x|z)P(y|z)}$$

$$X \perp Y | Z \quad \text{iff} \quad P(X, Y|Z) = P(X|Z) P(Y|Z) \quad \text{iff} \quad I(X, Y|Z) = 0$$

# Independence Tests and the Constraint-Based Approach

- Using the data  $D$

- Empirical distribution: 
$$\hat{P}(x_i, x_j) = \frac{\text{Count}(x_i, x_j)}{m}$$

- Mutual information: 
$$\hat{I}(X_i, X_j) = \sum_{x_i, x_j} \hat{P}(x_i, x_j) \log \frac{\hat{P}(x_i, x_j)}{\hat{P}(x_i)\hat{P}(x_j)}$$

- Similarly for conditional MI

- Use learning PDAG algorithm:

When algorithm asks:  $(X \perp Y | \mathbf{U})$  ?

- Use  $I(X, Y | \mathbf{U}) = 0$  ?

- No... doesn't happen

- Use  $I(X, Y | \mathbf{U}) < t$  for some  $t > 0$  ?

- ... based on some statistical test "t s.t.  $p < 0.05$ "

- Many other types of independence tests ...



# Independence Tests – II

- For discrete data:  $\chi^2$  statistic
  - measures how far the counts are from what we would expect given independence:

$$d_{\chi^2}(D) = \sum_{x,y} \frac{(O_{x,y} - E_{x,y})^2}{E_{x,y}} = \sum_{x,y} \frac{(N(x,y) - NP(x)P(y))^2}{NP(x)P(y)}$$

- p-value requires summing over all datasets of size N:

$$p(t) = P(\{D : d(D) > t\} \mid H_0, N)$$

- Expensive...  $\Rightarrow$  approximation
  - consider the expected distribution of  $d(D)$  (under the null hypothesis) as  $N \rightarrow \infty$
  - ... to define thresholds for a given significance





# Ex of Classical Hypothesis Testing

---

- Spin Belgian one-euro coin
  - $N = 250$ ... heads  $Y = 140$ ; tails 110.
- Distinguish two models,
  - $H_0$  = coin is unbiased (so  $p = 0.5$ )
  - $H_1$  = coin is biased  $p \neq 0.5$
- p-value is "less than 7%"
  - $p = P(Y \geq 140) + P(Y \leq 110) = 0.066$ :  
 $n=250; p = 0.5; y = 140$ ;  
 $p = (1 - \text{binocdf}(y-1, n, p)) + \text{binocdf}(n-y, n, p)$
- If  $Y = 141$ :  $p = 0.0497$   
 $\Rightarrow$  reject the null hypothesis at significance level 0.05.
- But is the coin really biased?



# build-PDAG Algorithm

---

build-PDAG can recover the true structure

- up to I-equivalence
- in  $O(N^3 2^d)$  time

if

- maximum number of parents over nodes is  $d$
- independence test oracle can handle  $\leq 2d + 2$  variables
- $\exists G =$  a I-map of  $P$ 
  - underlying distribution  $P$  is *faithful* to  $G$
  - $\neg \exists$  spurious independencies not sanctioned by  $G$
- Called IC or PC algorithm



# Eval of IC / PC alg

---

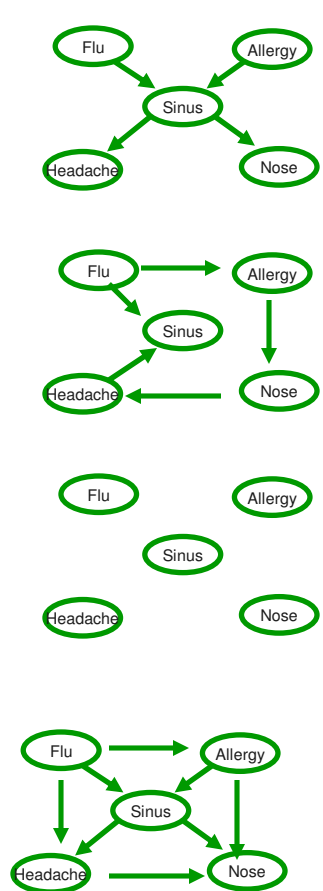
- Bad
  - Faithfulness assumption rules out certain CPDs
    - XOR.
  - Independence test typically unreliable
    - (especially given small data sets)
    - make many errors
  - One misleading independence test result can result in multiple errors in the resulting PDAG, so overall the approach is not robust to noise.
- Good
  - PC algorithm is less dumb than local search

# Score-based Approach



Possible DAG structures  
(gazillions)

Score of each Structure



$\langle x_1^{(1)}, \dots, x_n^{(1)} \rangle$   
...  
 $\langle x_1^{(m)}, \dots, x_n^{(m)} \rangle$

Learn Parameters  
+  
Evaluate ...

-15,000

**-10,000**

-20,000

-10,500

# Just use MLE parameters

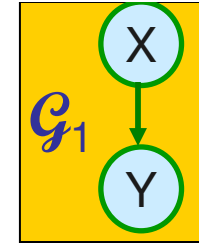
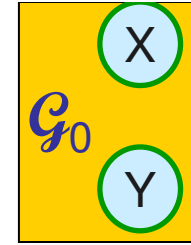
$$\begin{aligned} \blacksquare \max_{\mathcal{G}, \theta_{\mathcal{G}}} L( \langle \mathcal{G}, \theta_{\mathcal{G}} \rangle : \mathcal{D} ) &= \\ \max_{\mathcal{G}} \max_{\theta_{\mathcal{G}}} L( \langle \mathcal{G}, \theta_{\mathcal{G}} \rangle : \mathcal{D} ) &= \\ \max_{\mathcal{G}} L( \langle \mathcal{G}, \theta_{\mathcal{G}}^* \rangle : \mathcal{D} ) \end{aligned}$$

■ So...

seek the structure  $\mathcal{G}$  that achieves highest likelihood, given its MLE parameters  $\theta_{\mathcal{G}}^*$

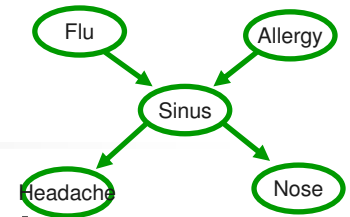
$$\blacksquare \text{Score}(\mathcal{G}, S) = \log L( \langle \mathcal{G}, \theta_{\mathcal{G}}^* \rangle : \mathcal{D} )$$

# Comparing Models



- $\mathcal{D} = \{ \langle x[1], y[1] \rangle, \dots, \langle x[M], y[M] \rangle \}$
- $\text{Score}(\mathcal{G}_0, \mathcal{D}) = \sum_m \log \theta_{x[m]}^* + \log \theta_{y[m]}^*$
- $\text{Score}(\mathcal{G}_1, \mathcal{D}) = \sum_m \log \theta_{x[m]}^* + \log \theta_{y[m] | x[m]}^*$
- $\text{Score}(\mathcal{G}_1, \mathcal{D}) - \text{Score}(\mathcal{G}_0, \mathcal{D})$ 
  - $= \sum_{x,y} M[x,y] \log \theta_{y[m]}^* - \sum_y M[y] \log \theta_{y[m]}^*$
  - $= M \sum_{x,y} p^*(x,y) \log[ p^*(y|x) / p(y) ]$
  - $= M I_{p^*}(X, Y)$
- $I_{p^*}(X, Y)$  = mutual information between  $X$  and  $Y$  in  $P^*$
- ... higher mutual info  $\Rightarrow$  stronger  $X \rightarrow Y$  dependency

# Information-theoretic interpretation of maximum likelihood

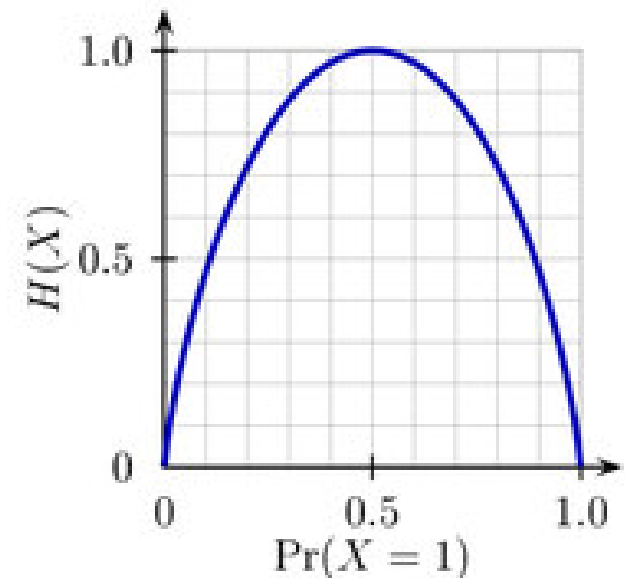


- Given structure  $\mathcal{G}$ , parameters  $\theta_{\mathcal{G}}$ , log likelihood of data  $\mathcal{D}$ :

$$\begin{aligned}
 \log P(\mathcal{D} \mid \theta_{\mathcal{G}}, \mathcal{G}) &= \sum_{j=1}^m \sum_{i=1}^n \log P \left( X_i = x_i^{(j)} \mid \mathbf{Pa}_{X_i} = \mathbf{x}^{(j)} [\mathbf{Pa}_{X_i}] \right) \\
 &= \sum_{i=1}^n \sum_{j=1}^m \log P \left( X_i = x_i^{(j)} \mid \mathbf{Pa}_{X_i} = \mathbf{x}^{(j)} [\mathbf{Pa}_{X_i}] \right) \\
 &= \sum_{i=1}^n \sum_{x_i, \mathbf{u}} \#(X_i = x_i, \mathbf{Pa}_{X_i} = \mathbf{u}) \log P \left( X_i = x_i \mid \mathbf{Pa}_{X_i} = \mathbf{u} \right) \\
 &= m \sum_{i=1}^n \sum_{x_i, \mathbf{u}} \frac{\#(X_i = x_i, \mathbf{Pa}_{X_i} = \mathbf{u})}{m} \log P \left( X_i = x_i \mid \mathbf{Pa}_{X_i} = \mathbf{u} \right) \\
 &= m \sum_{i=1}^n \sum_{x_i, \mathbf{u}} \hat{P}(X_i = x_i, \mathbf{Pa}_{X_i} = \mathbf{u}) \log P \left( X_i = x_i \mid \mathbf{Pa}_{X_i} = \mathbf{u} \right)
 \end{aligned}$$

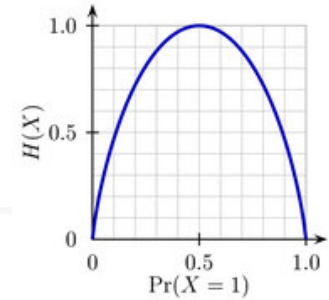
# Entropy

- Entropy of  $V = [p(V = 1), p(V = 0)]$  :  
$$H(V) = -\sum_{v_i} P(V = v_i) \log_2 P(V = v_i)$$
  
 $\equiv$  # of bits needed to obtain full info  
...average surprise of result of one "trial" of  $V$
- Entropy  $\approx$  measure of uncertainty





# Examples of Entropy



- Fair coin:

- $H(1/2, 1/2) = -1/2 \log_2(1/2) - 1/2 \log_2(1/2) = 1 \text{ bit}$
- ie, need 1 bit to convey the outcome of coin flip)

- Biased coin:

$$H(1/100, 99/100) = -1/100 \log_2(1/100) - 99/100 \log_2(99/100) = 0.08 \text{ bit}$$

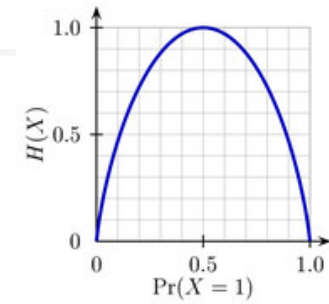
- As  $P(\text{heads}) \mapsto 1$ , info of actual outcome  $\mapsto 0$

$$H(0, 1) = H(1, 0) = 0 \text{ bits}$$

ie, no uncertainty left in source

$$(0 \times \log_2(0) = 0)$$

# Entropy & Conditional Entropy



## ■ Entropy of Distribution

- $H(X) = - \sum_i P(x_i) \log P(x_i)$
- “How ‘surprising’ variable is”
- Entropy = 0 when know everything... eg  $P(+x)=1.0$

## ■ Conditional Entropy $H(X | \mathbf{U})$ ...

- $H(X|\mathbf{U}) = - \sum_{\mathbf{u}} P(\mathbf{u}) \sum_i P(x_i|\mathbf{u}) \log P(x_i|\mathbf{u})$
- How much uncertainty is left in  $X$ , after observing  $\mathbf{U}$

$$H(X_i | \mathbf{Pa}_{X_i}) = - \sum_{x_i, \mathbf{u}} \hat{P}(X_i = x_i, \mathbf{Pa}_{X_i} = \mathbf{u}) \log P(X_i = x_i^{(j)} | \mathbf{Pa}_{X_i} = \mathbf{u})$$

# Information-theoretic interpretation of maximum likelihood ... 2

- Given structure  $\mathcal{G}$ , parameters  $\theta_{\mathcal{G}}$ , log likelihood of data  $\mathcal{D}$  is...

$$\begin{aligned} \uparrow \log \hat{P}(\mathcal{D} | \theta, \mathcal{G}) &= m \sum_i \sum_{x_i, \mathbf{u}} \hat{P}(x_i, \text{Pa}_{x_i, \mathcal{G}} = \mathbf{u}) \log \hat{P}(x_i | \text{Pa}_{x_i, \mathcal{G}} = \mathbf{u}) \\ &= m \sum_i -\hat{H}(X_i | \text{Pa}_{x_i, \mathcal{G}}) \\ &= -m \sum_i \hat{H}(X_i | \text{Pa}_{x_i, \mathcal{G}}) \quad \downarrow \end{aligned}$$

So  $\log P(\mathcal{D} | \theta, \mathcal{G})$  is LARGEST

when each  $H(X_i | \text{Pa}_{x_i, \mathcal{G}})$  is SMALL...

...ie, when parents of  $X_i$  are very INFORMATIVE about  $X_i$  !

# Score for Belief Network

- $\mathcal{I}(X, U) = H(X) - H(X | U)$   
 $\Rightarrow H(X | \text{Pa}_{X, \mathcal{G}}) = H(X) - \mathcal{I}(X, \text{Pa}_{X, \mathcal{G}})$

Doesn't involve the structure,  $\mathcal{G}$ !

- Log data likelihood

$$\log \hat{P}(\mathcal{D} | \theta, \mathcal{G}) = m \sum_i \hat{I}(X_i, \text{Pa}_{X_i, \mathcal{G}}) - m \sum_i \hat{H}(X_i)$$

- So use score:  $\sum_i \mathcal{I}(X_i, \text{Pa}_{X_i, \mathcal{G}})$



# Decomposable Score

---

- Log data likelihood

$$\log \hat{P}(\mathcal{D} \mid \theta, \mathcal{G}) = m \sum_i \hat{I}(X_i, \text{Pa}_{X_i, \mathcal{G}}) - m \sum_i \hat{H}(X_i)$$

- ... or perhaps just score:  $\sum_i \mathcal{J}(X_i, \text{Pa}_{X_i, \mathcal{G}})$

- Decomposable score:

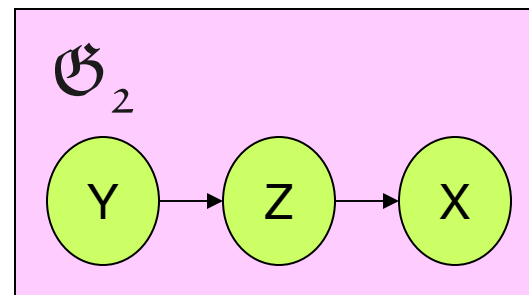
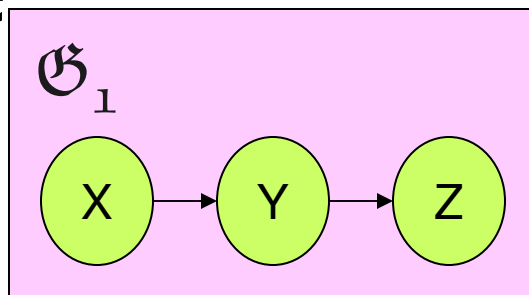
- Decomposes over families in BN (node and its parents)
- Will lead to significant computational efficiency!
- $\text{Score}(\mathcal{G} : \mathcal{D}) = \sum_i \text{FamScore}(X_i \mid \text{Pa}_{X_i} : \mathcal{D})$

# Using DeComposability

$$\log \hat{P}(\mathcal{D} \mid \theta, \mathcal{G}) = m \sum \hat{I}(x_i, \text{Pa}_{x_i, \mathcal{G}}) - m \sum \hat{H}(X_i)$$

$$\mapsto \sum_i \mathcal{J}(X_i, \text{Pa}_{X_i, \mathcal{G}}) + c$$

## ■ Compare



$$\blacksquare \mathcal{G}_1: \sum_i \mathcal{J}(X_i, \text{Pa}_{X_i, \mathcal{G}_1}) = \mathcal{J}(X, \{\}) + \mathcal{J}(Y, X) + \mathcal{J}(Z, Y)$$

$$= \mathcal{J}(Y, X) + \mathcal{J}(Z, Y)$$

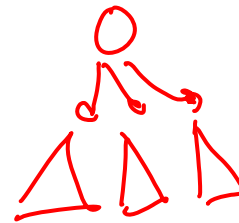
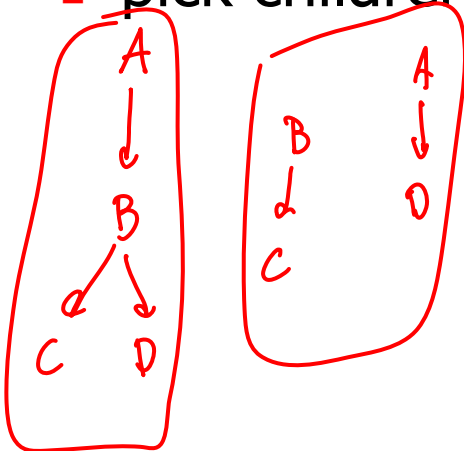
$$\blacksquare \mathcal{G}_2: \sum_i \mathcal{J}(X_i, \text{Pa}_{X_i, \mathcal{G}_2}) = \mathcal{J}(Y, \{\}) + \mathcal{J}(Z, Y) + \mathcal{J}(X, Z)$$

$$= \mathcal{J}(Z, Y) + \mathcal{J}(X, Z)$$

$$\blacksquare \dots \text{ so diff is } \mathcal{J}(Y, X) - \mathcal{J}(X, Z)$$

# How many trees are there?

- Tree:
  - $\exists$  one path between any two nodes (in skeleton)
  - Most nodes have 1 parent (+ root with 0 parents)
- How many:
  - One: pick root
  - pick children ... for each child ... another tree



$$\sim 2^{\Theta(n \lg n)}$$

**Nonetheless...  $\exists$  efficient optimal alg to find OPTIMAL tree**



# Best Tree Structure

---

$$\log \hat{P}(\mathcal{D} \mid \theta, \mathcal{G}) = m \sum_i \hat{I}(x_i, \text{Pa}_{x_i, \mathcal{G}}) - m \sum_i \hat{H}(X_i)$$

- Identify tree with set  $\mathcal{F} = \{ \text{Pa}(X) \}$ 
  - each  $\text{Pa}(X)$  is  $\{\}$ , or another variable
- Optimal tree, given data, is
$$\text{argmax}_{\mathcal{F}} m \sum_i I( X_i, \text{Pa}(X_i) ) - m \sum_i H(X_i)$$
$$= \text{argmax}_{\mathcal{F}} \sum_i I( X_i, \text{Pa}(X_i) )$$
  - ... as  $\sum_i H(X_i)$  does not depend on structure
- So ... want parents  $\mathcal{F}$  s.t.
  - tree structure
  - maximizes  $\sum_i I( X_i, \text{Pa}(X_i) )$



# Chow-Liu Tree Learning Alg

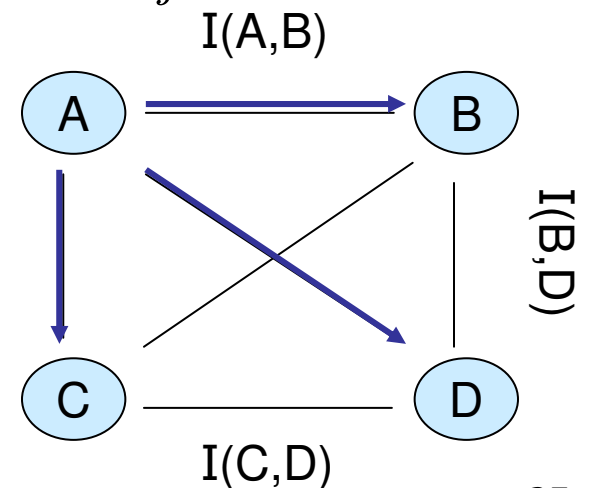
- For each pair of variables  $X_i, X_j$ 
  - Compute empirical distribution:

$$\hat{P}(x_i, x_j) = \frac{\text{Count}(x_i, x_j)}{m}$$

- Compute mutual information:

$$\hat{I}(X_i, X_j) = \sum_{x_i, x_j} \hat{P}(x_i, x_j) \log \frac{\hat{P}(x_i, x_j)}{\hat{P}(x_i)\hat{P}(x_j)}$$

- Define a graph
  - Nodes  $X_1, \dots, X_n$
  - Edge (i,j) gets weight  $\hat{I}(X_i, X_j)$
- Find Maximal Spanning Tree
- Pick a node for root, dangle...



# Chow-Liu Tree Learning Alg ... 2

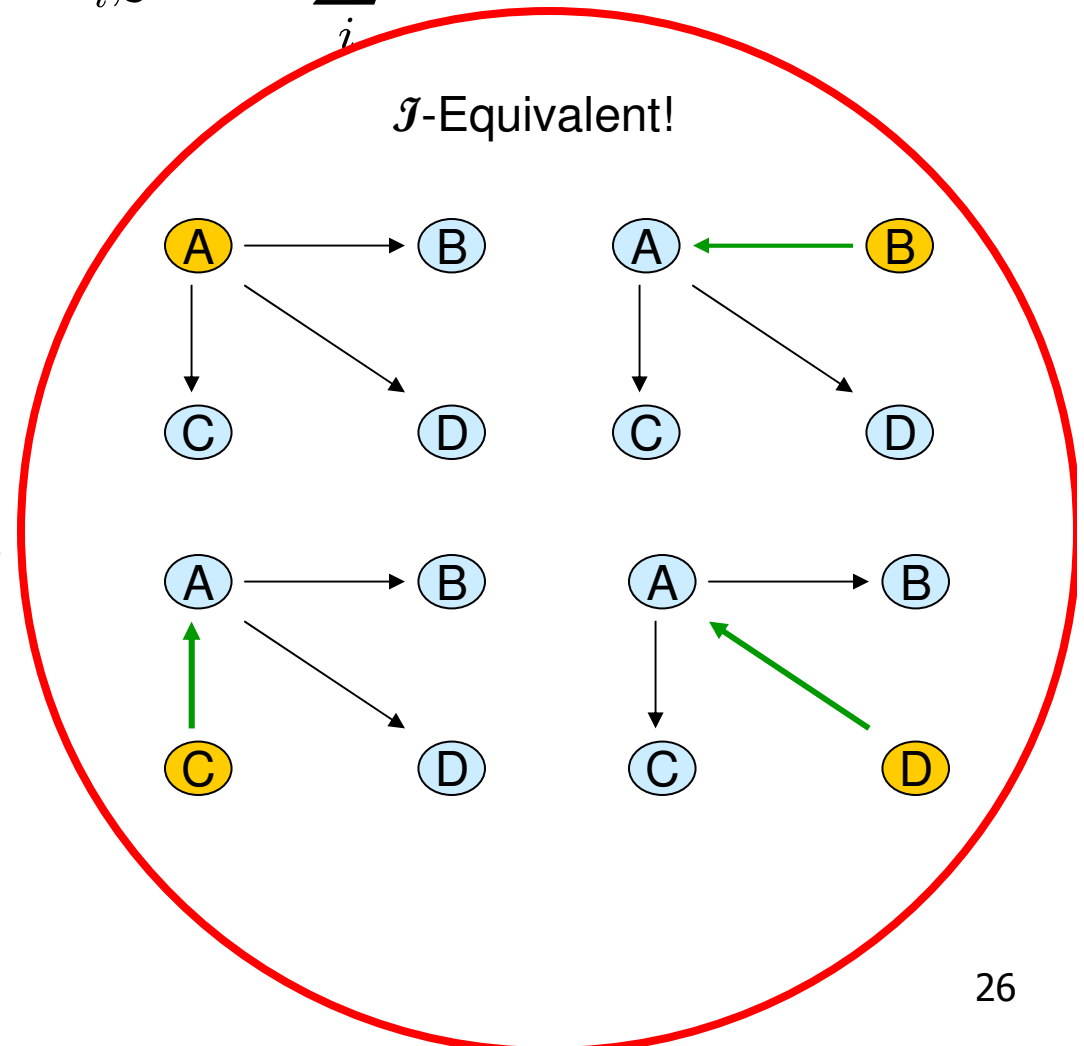
$$\log \hat{P}(\mathcal{D} \mid \theta, \mathcal{G}) = m \sum_i \hat{I}(x_i, \text{Pa}_{x_i, \mathcal{G}}) - m \sum_i \hat{H}(X_i)$$

## ■ Optimal tree BN

- ...
- Compute maximum weight spanning tree
- Directions in BN:
  - pick any node as root, ...doesn't matter which!
  - breadth-first-search defines directions

## ■ Score Equivalence:

If  $\mathcal{G}$  and  $\mathcal{G}'$  are  $\mathcal{I}$ -equiv, then scores are same





# Chow-Liu (CL) Results

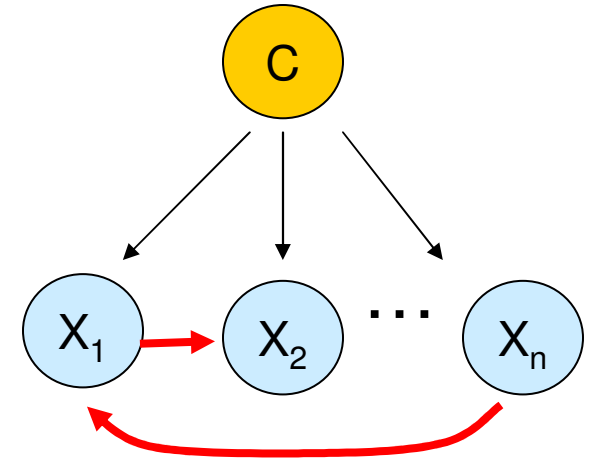
---

- If distribution  $P$  is tree-structured, CL finds CORRECT one
- If distribution  $P$  is NOT tree-structured, CL finds tree structured  $Q$  that has min'l KL-divergence –  $\operatorname{argmin}_Q \text{KL}(P; Q)$
- Even though  $2^{\theta(n \log n)}$  trees, CL finds BEST one in poly time  $O(n^2 [m + \log n])$

# Using Chow-Liu to Improve NB

- Naïve Bayes model

- $X_i \perp X_j \mid C$
- Ignores correlation between features
- What if  $X_1 = X_2$ ? **Double count...**



- Avoid by conditioning features on one another

- Tree Augmented Naïve bayes (TAN)

[Friedman et al. '97]

$$\hat{I}(X_i, X_j \mid C) = \sum_{c, x_i, x_j} \hat{P}(c, x_i, x_j) \log \frac{\hat{P}(x_i, x_j \mid c)}{\hat{P}(x_i \mid c) \hat{P}(x_j \mid c)}$$

All but ONE feature have 2 parents: C,  $X_i$



# Can we extend Chow-Liu ?

---

- (Approximately learning) models with tree-width up to  $k$ 
  - [Narasimhan & Bilmes '04]
  - But,  $O(n^{k+1})$ ...
    - and more subtleties



# Learning BN structures... so far

---

- Decomposable scores
  - Maximum likelihood
  - Information theoretic interpretation
- Best tree (Chow-Liu)
- Best TAN
- Nearly best k-treewidth (in  $O(N^{k+1})$ )

# Maximum likelihood score overfits!

$$\log \hat{P}(\mathcal{D} | \theta, \mathcal{G}) = m \sum_i \hat{I}(X_i, \text{Pa}_{X_i, \mathcal{G}}) - m \sum_i \hat{H}(X_i)$$

- Adding a parent never decreases score!!!

- *Facts:*  $H(X | \text{Pa}_{X, \mathcal{G}}) = H(X) - I(X, \text{Pa}_{X, \mathcal{G}})$

$$H(X | A) \geq H(X | A \cup Y)$$

- $I(X_i, \text{Pa}_{X_i, \mathcal{G}} \cup Y) = H(X_i) - H(X_i | \text{Pa}_{X_i, \mathcal{G}} \cup Y)$   
 $\geq H(X_i) - H(X_i | \text{Pa}_{X_i, \mathcal{G}})$   
 $= I(X_i, \text{Pa}_{X_i, \mathcal{G}})$

- So score increases as we add edges!

- Best is COMPLETE Graph
- ... overfit !

# How to Evaluate a Model?

## Training Data

SNP1	SNP2	SNP3	...	SNP53	Bleed?
G/A	C/C	T/T	...	T/C	No
A/A	C/C	A/T	...	T/T	Yes
A/A	C/T	A/A	...	T/T	Yes
:	:	:		:	:
G/A	C/T	A/A	...	T/T	No

Training Set Error  
... too optimistic

## TRAIN

SNP1	SNP2	SNP3	...	SNP53	Bleeding?
C/G	A/G	T/T	...	T/C	No
T/C	C/C	A/A	...	T/T	Yes
:	:	:		:	:
G/A	T/C	G/G	...	T/C	No

## TEST

SNP1	SNP2	SNP3	...	SNP53
G/A	C/C	T/T	...	T/C
A/A	C/C	A/T	...	T/T
:	:	:		:
G/A	C/T	A/A	...	T/T

Find Param

Compute Prob

-38



# How to Evaluate a Model?

## Training Data

SNP1	SNP2	SNP3	...	SNP53	Bleed?
G/A	C/C	T/T	...	T/C	No
A/A	C/C	A/T	...	T/T	Yes
A/A	C/T	A/A	...	T/T	Yes
:	:	:		:	:
G/A	C/T	A/A	...	T/T	No

## TRAIN

SNP1	SNP2	SNP3	...	SNP53	Bleeding?
C/G	A/G	T/T	...	T/C	No
T/C	C/C	A/A	...	T/T	Yes
:	:	:		:	:
G/A	T/C	G/G	...	T/C	No

## TEST

SNP1	SNP2	SNP3	...	SNP53
G/A	C/C	T/T	...	T/C
A/A	C/C	A/T	...	T/T
:	:	:		:
G/A	C/T	A/A	...	T/T

Find Param

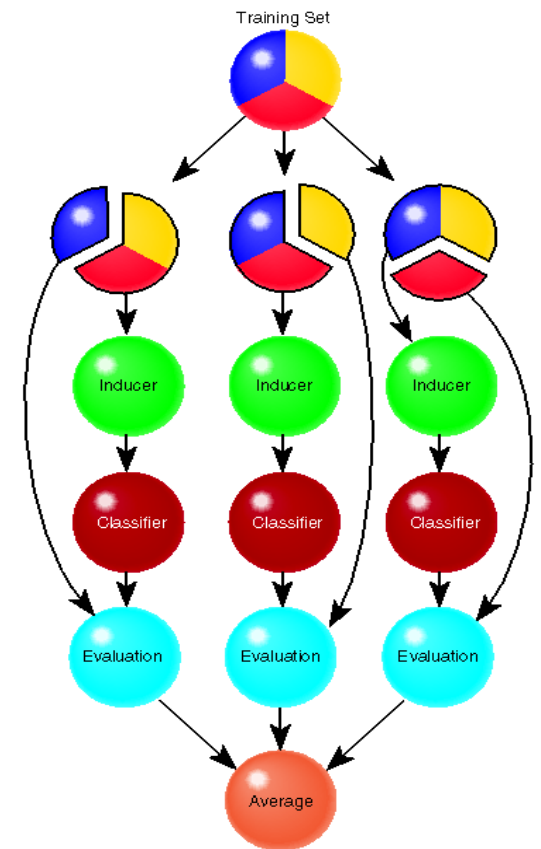
Compute Prob

-53

Simple Hold-out Set Error  
... slightly pessimistic

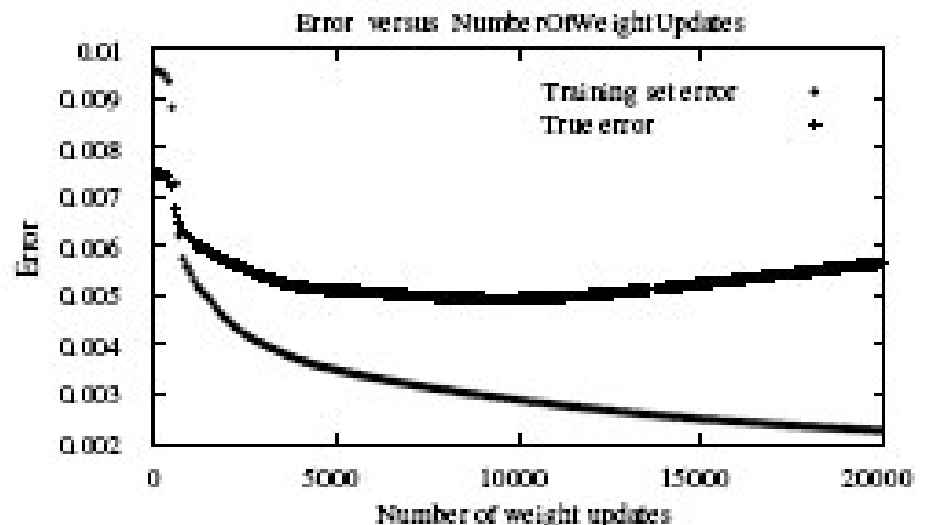
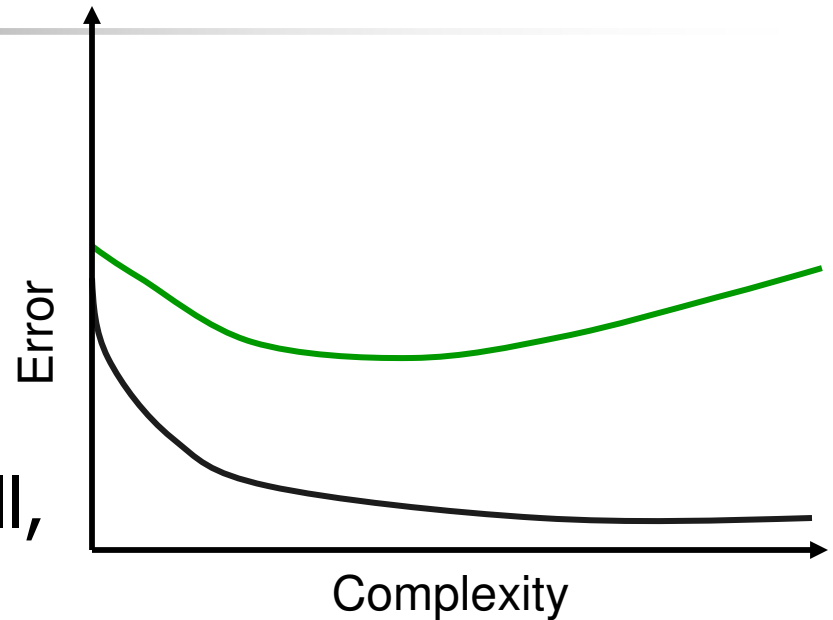
# How to Evaluate a Model ?

- K-fold Cross Validation
  - Eg, K=3
- Not as pessimistic
  - every point is test example, once



# Overfitting

- So far:  
Find parameters/structure that "fit" the training data
- If too many parameters, will match TRAINING data well, but NOT new instances
- **Overfitting!**
- Regularizing, Bayesian approach, ...



# Bayesian Score

- Prior distributions:

- Over structures
- Over parameters of a structure

Goal: Prefer simpler structures... regularization ...

- Posterior over structures given data:

- $P(\mathcal{G}|\mathcal{D}) \propto P(\mathcal{D}|\mathcal{G}) \times P(\mathcal{G})$

Posterior

Likelihood

Prior over Graphs

Prior over Parameters

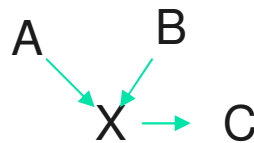
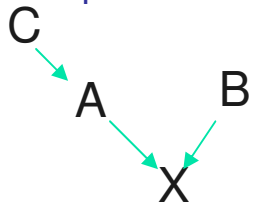
- $P(\mathcal{D}|\mathcal{G}) = \int_{\Theta} P(\mathcal{D} | \mathcal{G}, \Theta) P(\Theta|\mathcal{G}) d\Theta$

$$\log P(\mathcal{G} | D) \approx \log P(\mathcal{G}) + \log \int_{\theta_{\mathcal{G}}} P(D | \mathcal{G}, \theta_{\mathcal{G}}) P(\theta_{\mathcal{G}}|\mathcal{G}) d\theta_{\mathcal{G}}$$

# Towards a decomposable Bayesian score

$$\log P(\mathcal{G} | D) \approx \log P(\mathcal{G}) + \log \int_{\theta_{\mathcal{G}}} P(D | \mathcal{G}, \theta_{\mathcal{G}}) P(\theta_{\mathcal{G}} | \mathcal{G}) d\theta_{\mathcal{G}}$$

- **Local and global parameter independence**  $\theta_{Y|+X} \perp \theta_X$
- Prior satisfies **parameter modularity**:
  - If  $X_i$  has same parents in  $\mathcal{G}$  and  $\mathcal{G}'$ , then parameters have same prior



$\Theta(X; A, B)$  same in both structures

- Structure prior  $P(\mathcal{G})$  satisfies **structure modularity**
  - Product of terms over families
  - Eg,  $P(\mathcal{G}) \propto c^{|\mathcal{G}|}$   $|\mathcal{G}| = \#edges; c < 1$
- ... then ... Bayesian score decomposes along families!
  - $\log P(\mathcal{G} | \mathcal{D}) = \sum_X \text{ScoreFam}(X | Pa_X : \mathcal{D})$

# Marginal Posterior

- Given  $\theta \sim \text{Beta}(1,1)$ , what is probability of  $\langle H, T, T, H, H \rangle$ ?

$$\begin{aligned}
 & P(f_1=H, f_2=T, f_3=T, f_4=H, f_5=H \mid \theta \sim \text{Beta}(1,1)) \\
 &= P(f_1=H \mid \theta \sim \text{Beta}(1,1)) \times \\
 &\quad P(f_2=T, f_3=T, f_4=H, f_5=H \mid f_1=H, \theta \sim \text{Beta}(1,1)) \\
 &= \frac{1}{2} \times P(f_2=T, f_3=T, f_4=H, f_5=H \mid \theta \sim \text{Beta}(2,1)) \\
 &= \frac{1}{2} \times P(f_2=T \mid \theta \sim \text{Beta}(2,1)) \times \\
 &\quad P(f_3=T, f_4=H, f_5=H \mid f_2=T, \theta \sim \text{Beta}(2,1)) \\
 &= \frac{1}{2} \times \frac{1}{3} \times P(f_3=T, f_4=H, f_5=H \mid \theta \sim \text{Beta}(2,2)) \\
 &= \frac{1}{2} \times \frac{1}{3} \times \frac{2}{4} \times \frac{2}{5} \times P(f_5=H \mid \theta \sim \text{Beta}(2,3)) \\
 &= \frac{1}{2} \times \frac{1}{3} \times \frac{2}{4} \times \frac{2}{5} \times \frac{3}{6} \\
 &= (1 \times 2 \times 3) \times (1 \times 2) / (2 \times 3 \times 4 \times 5)
 \end{aligned}$$

# Marginal Posterior... con't

- Given  $\theta \sim \text{Beta}(a,b)$ , what is  $P[ \langle H, T, T, H, H \rangle ]$  ?
- $P( f_1=H, f_2=T, f_3=T, f_4=H, f_5=H \mid \theta \sim \text{Beta}(a,b) )$   
 $= P( f_1=H \mid \theta \sim \text{Beta}(a,b) ) \times$   
 $P( f_2=T, f_3=T, f_4=H, f_5=H \mid f_1=H, \theta \sim \text{Beta}(a,b) )$   
 $= a/(a+b) \times P( f_2=T, f_3=T, f_4=H, f_5=H \mid \theta \sim \text{Beta}(a+1,b) )$

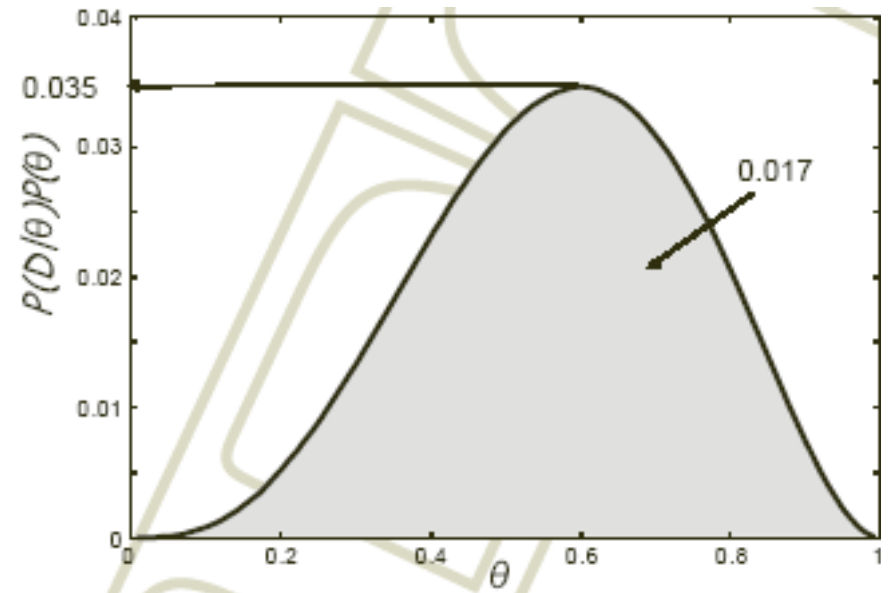
$$= \frac{a}{a+b} \frac{b}{a+b+1} \frac{b+1}{a+b+2} \frac{a+1}{a+b+3} \frac{a+2}{a+b+4}$$

$$= \frac{a \times (a+1) \times (a+2) \times b \times (b+1)}{(a+b)(a+b+1)(a+b+2)(a+b+3)(a+b+4)}$$

$$= \frac{\Gamma(a+m_H)}{\Gamma(a)} \frac{\Gamma(b+m_T)}{\Gamma(b)} \frac{\Gamma(a+b)}{\Gamma(a+b+m)}$$

# Marginal, vs Maximal, Likelihood

- Data  $\mathcal{D} = \langle H, T, T, H, H \rangle$
- $\theta^* = \operatorname{argmax}_{\theta} P(\mathcal{D} | \theta) = 3/5$ 
  - ... Here:  $P(\mathcal{D} | \theta^*) = (3/5)^3 (2/5)^2 \approx 0.035$
  - Or Bayesian,  
from Beta(1,1),  $\theta_{B(1,1)}^* = 4/7$
- Marginal
  - $\prod_i P(x_i | x_1, \dots, x_{i-1})$
  - kinda like cross validation:  
Evaluate each instance,  
wrt previous instance







# Marginal Probability of Graph

$$\log P(D | \mathcal{G}) = \log \int_{\theta_{\mathcal{G}}} P(D | \mathcal{G}, \theta_{\mathcal{G}}) P(\theta_{\mathcal{G}} | \mathcal{G}) d\theta_{\mathcal{G}}$$

- Given complete data, independent parameters, ...

$$P(D|G) = \prod_i \prod_{u_i \in \text{Val}(P_{\alpha_{X_i}})} \frac{\Gamma(\alpha_{X_i|u_i}^G)}{\Gamma(\alpha_{X_i|u_i}^G + M[u_i])} \prod_{x_i^j \in \text{Val}(X_i)} \frac{\Gamma(\alpha_{x_i^j|u_i}^G + M[x_i^j, u_i])}{\Gamma(\alpha_{x_i^j|u_i}^G)}$$

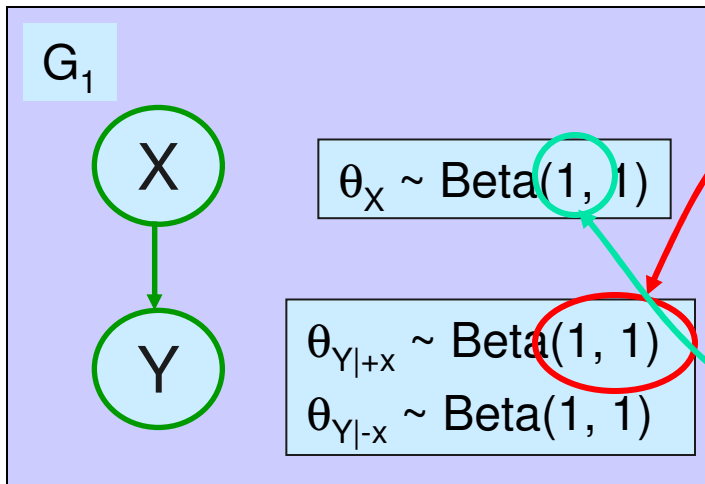


# Priors for General Graphs

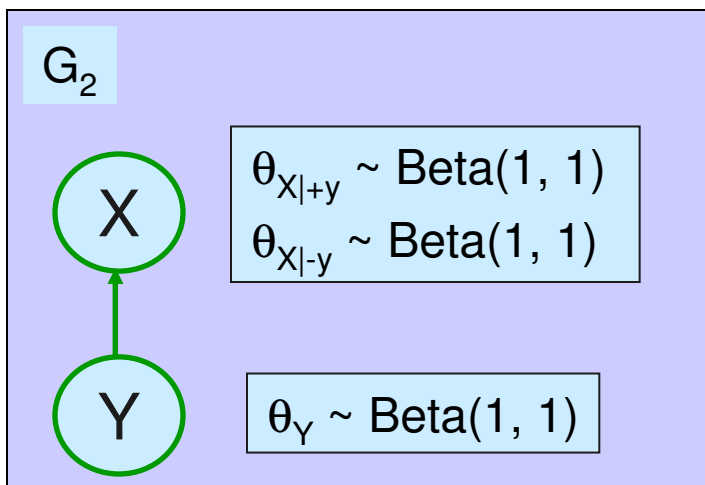
---

- For finite datasets, prior is important!
- Prior over structure satisfying prior modularity
  - Eg,  $P(\mathcal{G}) \propto c^{|\mathcal{G}|}$   $|\mathcal{G}| = \# \text{edges}; c < 1$
- What is good prior over *all* parameters?
  - *K2 prior*: fix  $\alpha \in \mathbb{R}^+$ , set  $\theta_{X_i | \text{Pa} X_i} \sim \text{Dirichlet}(\alpha, \dots, \alpha)$
  - Effective sample size, wrt  $X_i$ ?
    - If 0 parents:  $k \times \alpha$
    - If 1 binary parent:  $2 k \times \alpha$
    - If  $d$   $k$ -ary parents:  $k^d k \times \alpha$
  - So  $X_i$  "effective sample size" depends on #parental assignments
    - More parents  $\Rightarrow$  strong prior... doesn't make sense!
  - K2 is "inconsistent"

# Priors for Parameters

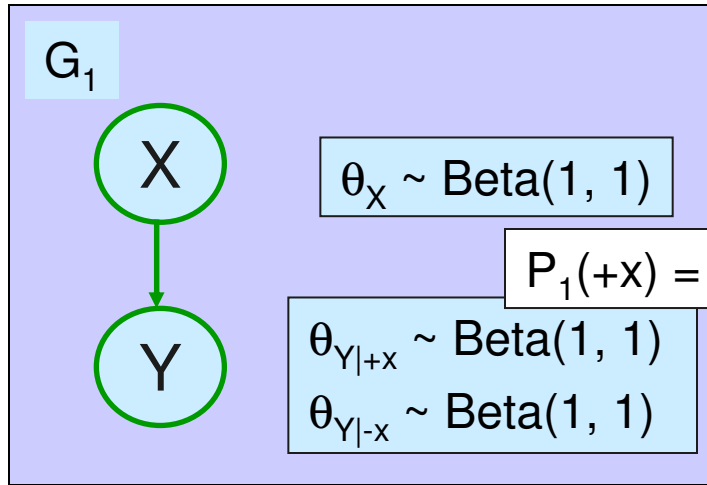


- Does this make sense?
  - $\text{EffectiveSampleSize}(\theta_{Y|+x}) = 2$
  - But only 1 example  $\sim$  "+x" ??

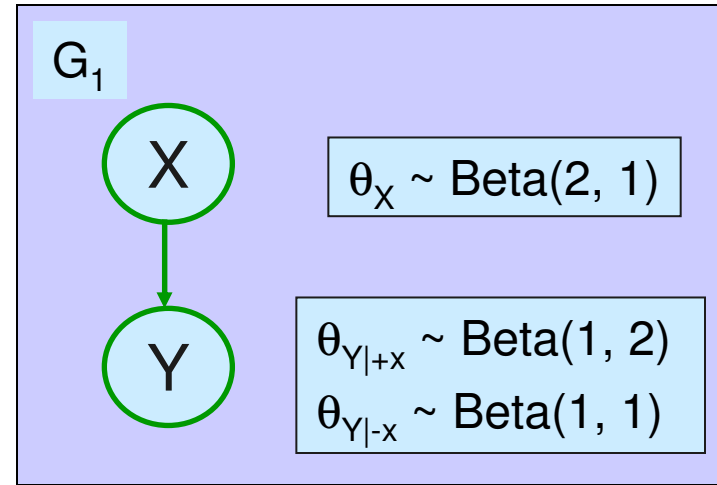


- $\mathcal{I}$ -Equivalent structure
- What happens after  $[+x, -y]$ ?
  - Should be the same!!

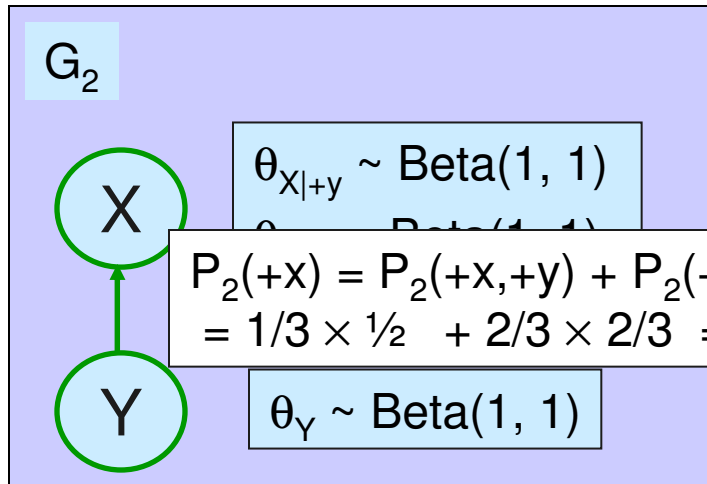
# Priors for Parameters



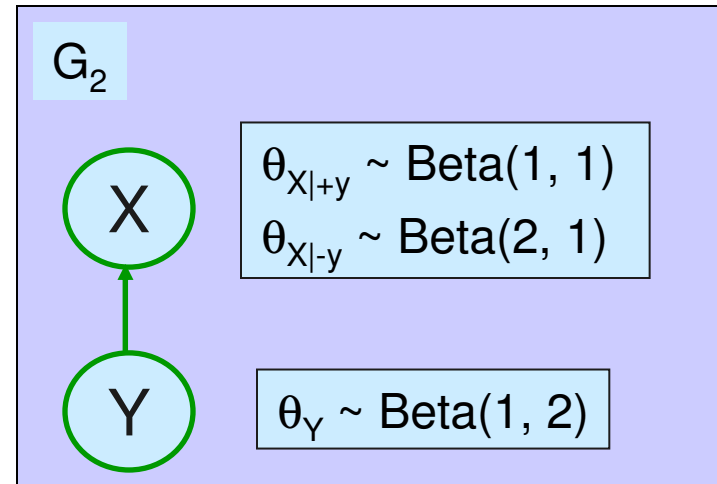
$P_1(+x) = 2/3$



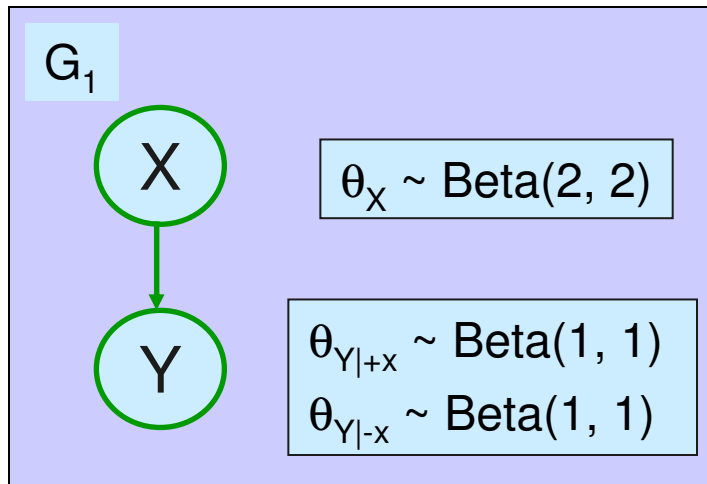
[+X, -y]



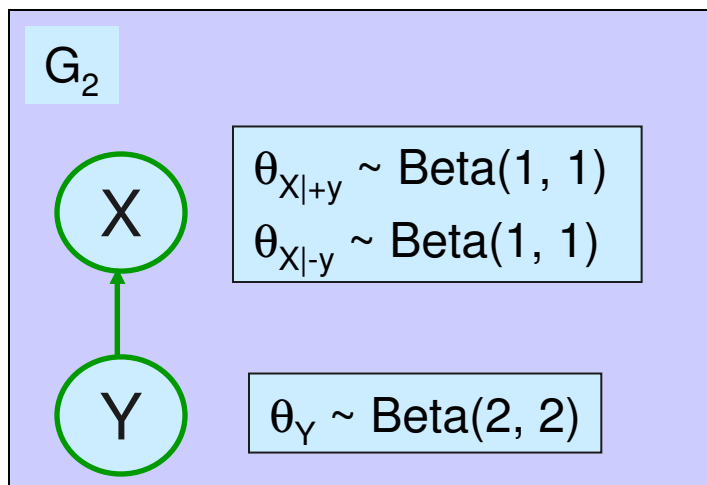
$P_2(+x) = P_2(+x, +y) + P_2(+x, -y)$   
 $= 1/3 \times 1/2 + 2/3 \times 2/3 = 11/18 !!!$



# BDe Priors

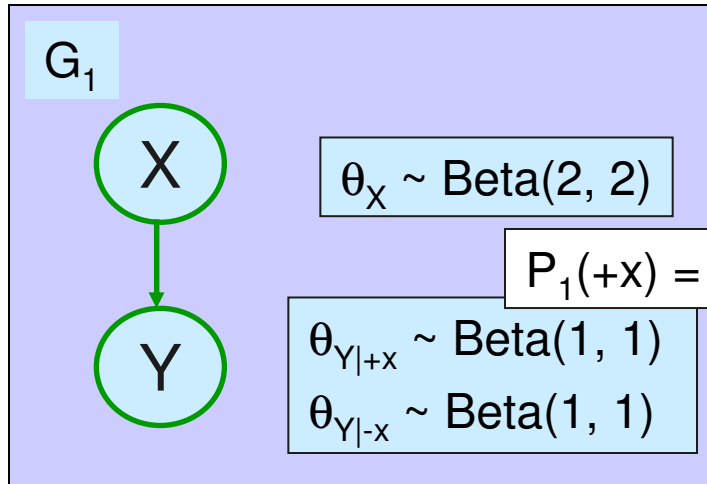


- This makes more sense:
  - $\text{EffectiveSampleSize}(\theta_{Y|+x}) = 2$
  - Now  $\approx \exists$  2 examples  $\sim$  “+x” ??

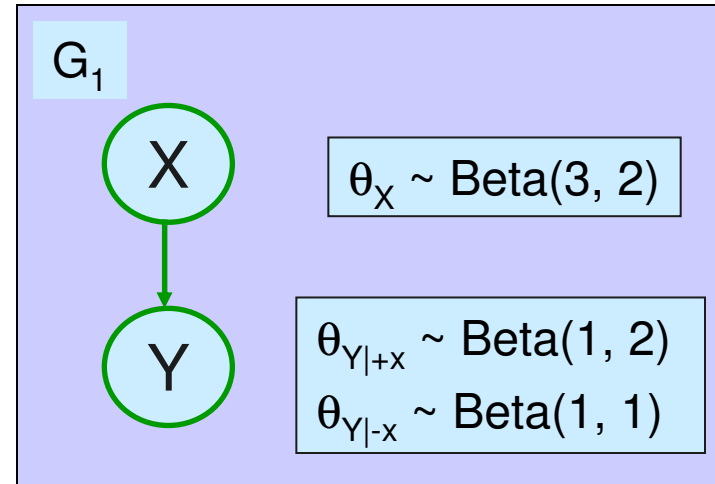


- $\mathcal{I}$ -Equivalent structure
- Now what happens after [+x, -y] ?

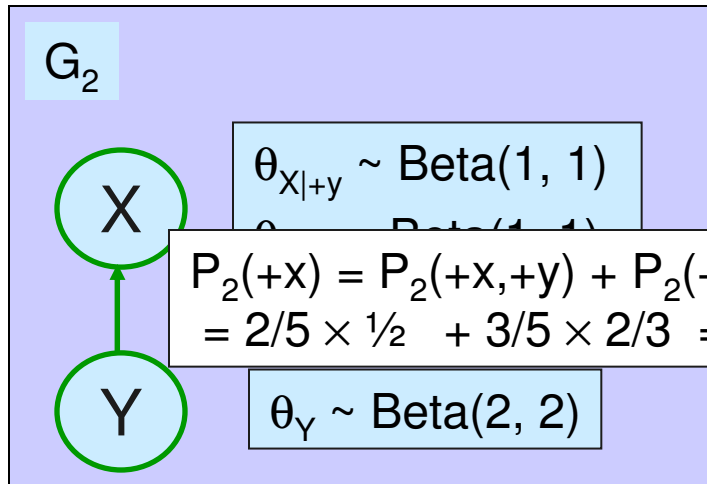
# BDe Priors



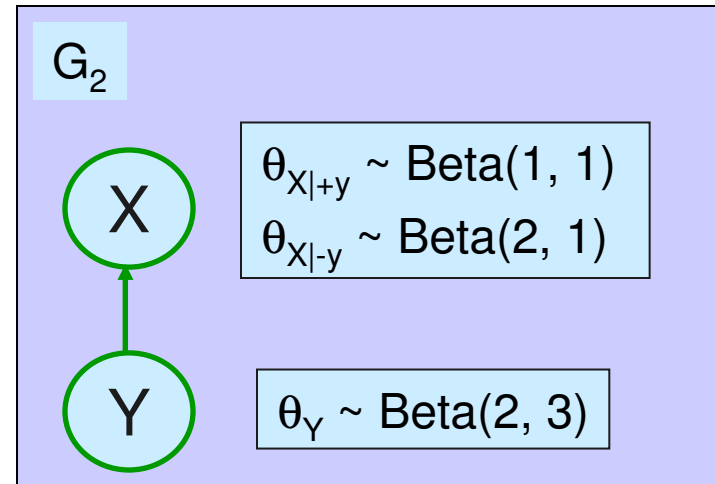
$P_1(+x) = 3/5$



[+X, -y]



$P_2(+x) = P_2(+x, +y) + P_2(+x, -y)$   
 $= 2/5 \times 1/2 + 3/5 \times 2/3 = 3/5 !!!$





# BDe Prior

---

- View Dirichlet parameters as “fictitious samples”
  - equivalent sample size
- Pick a fictitious sample size  $m'$
- For each possible family, define a prior distribution  $P(X_i, \mathbf{Pa}_{X_i})$ 
  - Represent with a BN
  - Usually independent (product of marginals)
    - $P(X_i, \mathbf{Pa}_{X_i}) = P'(x_i) \prod_{x_j \in \mathbf{Pa}[X_i]} P'(x_j)$
    - $P(\theta[x_i | \mathbf{Pa}_{X_i} = u]) = \text{Dir}(m' P'(x_i=1, \mathbf{Pa}_{X_i} = u), \dots, m' P'(x_i=k, \mathbf{Pa}_{X_i} = u))$
    - Typically,  $P'(X_i) = \text{uniform}$



# Summary wrt Learning BN Structure

---

- Decomposable scores
  - Data likelihood
  - Information theoretic interpretation
  - Bayesian
    - ┌ BIC approximation
- Priors
  - Structure and parameter assumptions
  - BDe if and only if score equivalence
- Best tree (Chow-Liu)
- Best TAN
  - ┌ Nearly best k-treewidth (in  $O(N^{k+1})$ )
  - ┌ Search techniques
    - ┌ Search through orders
    - ┌ Search through structures
  - ┌ Bayesian model averaging