Outline

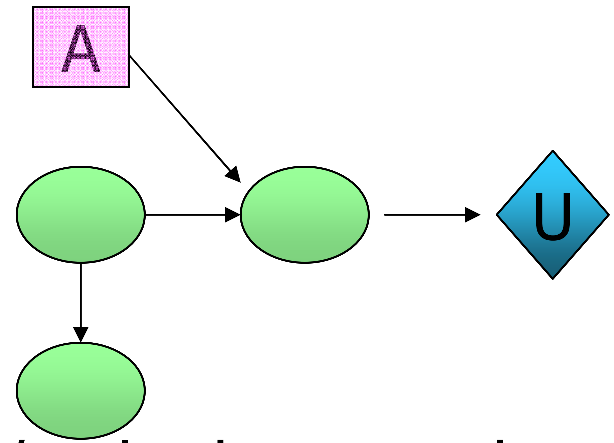
- Motivation
- What is a Belief Net?
  - Example
  - Inference
    - Maximize Expected Utility
  - Semantics
  - Relation to other Models
- Learning a Belief Net

- My Research

Utility-Based Agents

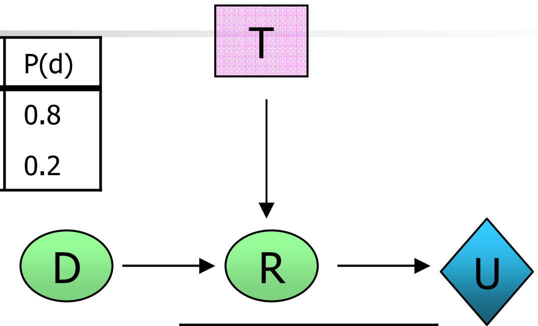
- MEU Principle:
  Agent should act to maximize expected utility
- Choose action $\boxed{A^* = \text{argmax}_A \{ EU(A|O)\}}$
  that maximizes
  expected utility of state after A,
  given prior observations O:
  $EU(A \mid O) =$
  $= \sum_{S'} P(S'|A,O)\, U(S')$
  $= \sum_{S'} \sum_S P(S \mid O)\, P(S' \mid S,A)\, U(S')$
  $= \sum_{S'} \sum_S [\alpha\, P(O \mid S)\, P(S)]\;\; P(S' \mid S,A)\, U(S')$
- Given simple assumptions, this is best possible action!
  (Average of utility, not of ~~utility²~~, not ~~minimaxing…~~)
- Good decision, bad outcome.

Decision Network



- Chance Nodes: S, O, S′
  - Bayesian net ≡ decision diagram w/ only chance nodes
  - Specify: P(S), P(O | S), P(S′ | S, A)
  - Here: S ≡ Current State   O ≡ Observation
    S′ ≡ Resulting State
- Decision Nodes: A
  - represents decision/action to make.
  - Specify: set of possible actions $a \in$ Dom(A)
- Utility Node(s): U
  - represents utility of each value-set of its parent chance variables
  - Specify: set of U(s′) for each $s' \in$ Dom(S′)

Perform a Medical Treatment?



d	P(d)
0	0.8
1	0.2

d	t	P(+r|d,t)
0	0	.001
0	1	.001
1	0	.950
1	1	.010

r	U(r)
0	0
1	-1000

- $EU(T = 1) =$
  $\sum_r P(R = r \mid T = 1)\, U(R = r)$

  $EU(T = 0) =$
  $\sum_r P(R = r \mid T = 0)\, U(R = r)$

- $P(R = 1 \mid T = 1) =$
  $\sum_d P(R = 1, D = d \mid T = 1)$

  $= \sum_d P(R = 1 \mid D = d, T = 1)\, P(D = d)$

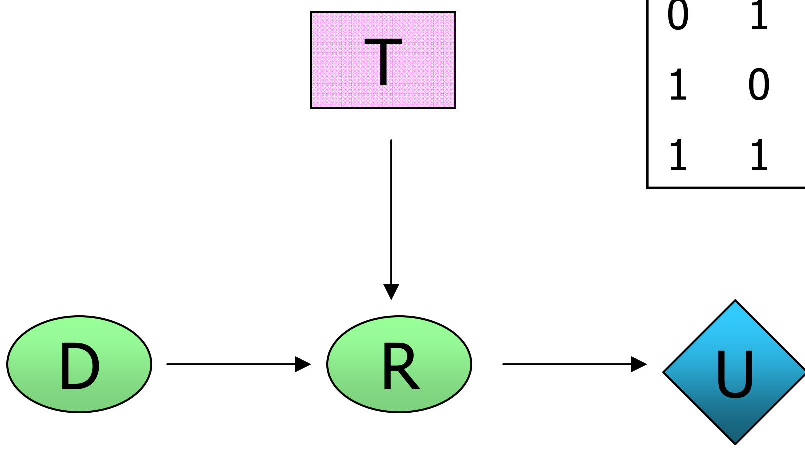
  $= P(R = 1 \mid D = 0, T = 1)\, P(D = 0) + P(R = 1 \mid D = 1, T = 1)\, P(D = 1)$
  $= (0.001 \times 0.8) + (0.01 \times 0.2) = 0.0028$

- $P(R = 0 \mid T = 1) = 1 - P(R = 1 \mid T = 1) = 0.9972$
- Similarly:
  - $P(R = 1 \mid T = 0) = 0.1908$
  - $P(R = 0 \mid T = 0) = 0.8092$

Medical Treatment (con't)

d	t	P(+r|d,t)
0	0	.001
0	1	.001
1	0	.950
1	1	.010

d	P(d)
0	0.8
1	0.2

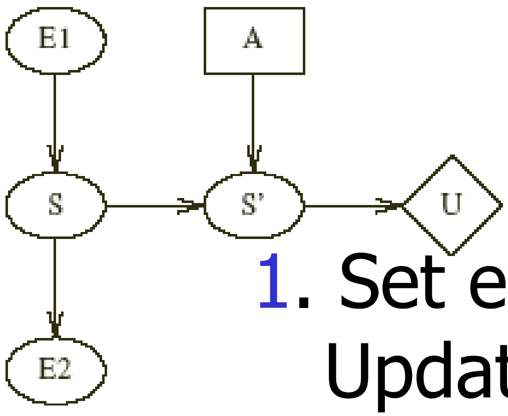


r	U(r)
0	0
1	-1000

	$P(R\|T)$		$U(R)$		
T	0	1	0	1	$EU(T)$
0	.8092	.1908	0	-1000	-190.8
1	.9972	.0028	0	-1000	-2.8

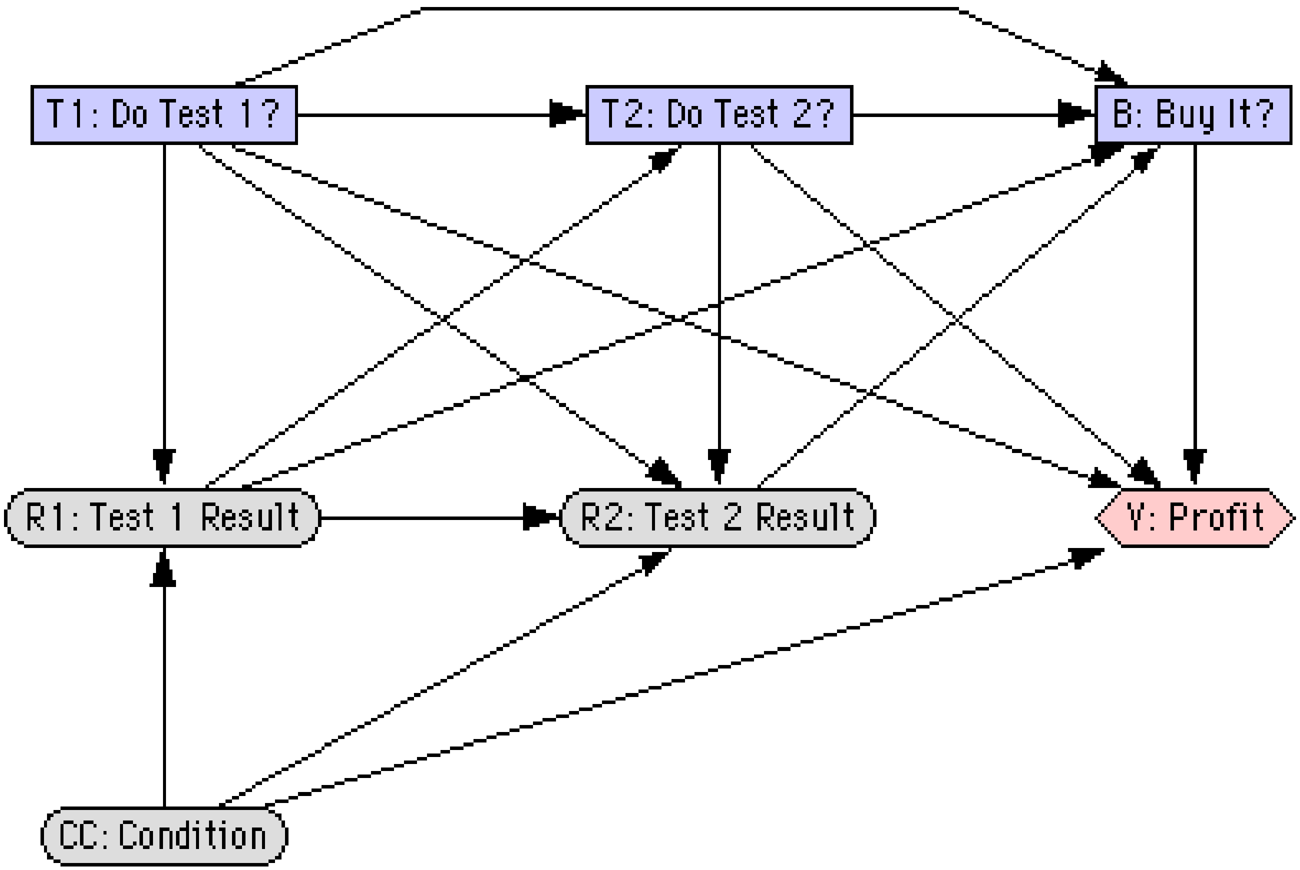
$\Leftarrow$ chosen action

Evaluating a Decision Network



1. Set evidence variables E_1, E_2
   Update distribution over current state S
2. For each possible action a of decision node A
   (a) Set decision node A to a
   (b) For each parent $\{ S' \}$ of utility node U:
       Calculate posterior probability of S
       Here, just $P(S' \mid E_1, E_2, A = a)$
   (c) Calculate expected utility for action a:
       $EU(A \mid E_1, E_2) = \sum_{S'} P(S' \mid E_1, E_2, a) \, U(S')$
3. Choose action $a^* = \arg\max_a \{ EU(a \mid \dots) \}$
   with highest expected utility

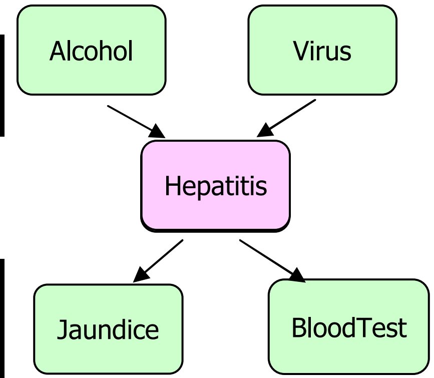
Decision Net: Test/Buy a Car



Outline

- Motivation
- What is a Belief Net?
  - Example
  - Inference
  - Semantics
    - d-separation
    - Noisy-Or
    - Continuous variables
  - Relation to other Models
- Learning a Belief Net
- My Research

Belief Nets



P(A=1)
0.32

P(V=1)
0.20

a	v	P(H=1|a ,v)
1	1	0.82
1	0	0.10
0	1	0.45
0	0	0.04

h	P(J=1| h)
1	0.8
0	0.3

h	P(B=1| h)
1	0.98
0	0.01

- DAG structure
  - Each node ≡ Variable *v*
  - *v* depends (only) on its parents

  \+ conditional prob: $P(v_i \mid \text{parent}_i = \langle 0,1,\ldots \rangle)$

- *v is INDEPENDENT* of non-descendants, given assignments to its parents

- Given H = 1,
  - A has no influence on J
  - J has no influence on B
  - etc.

Factoid: Chain Rule

- $P(A,B,C) = P(A \mid B,C)\, P(B,C)$

  $= P(A \mid B,C)\, P(B|C)\, P(C)$

- In general:

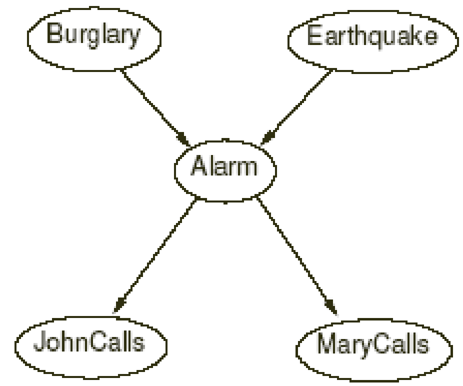
  $P(X_1,X_2, \ldots ,X_m) =$

  $P(X_1 \mid X_2, \ldots ,X_m)\, P(X_2, \ldots ,X_m) =$

  $P(X_1 \mid X_2, \ldots ,X_m)\, P(X_2 \mid X_3, \ldots ,X_m)\, P(X_3, \ldots ,X_m) =$

  $\prod_i P(X_i \mid X_{i+1}, \ldots ,X_m)$

Joint Distribution



Node is INDEPENDENT of non-descendants, given assignments to its parents

$$P(+j, +m, +a, -b, -e)$$

$$= \cancel{P(+j \mid +m, +a, -b, -e)} \xrightarrow{J \perp \{M,B,E\} \mid A} P(+j \mid +a)$$

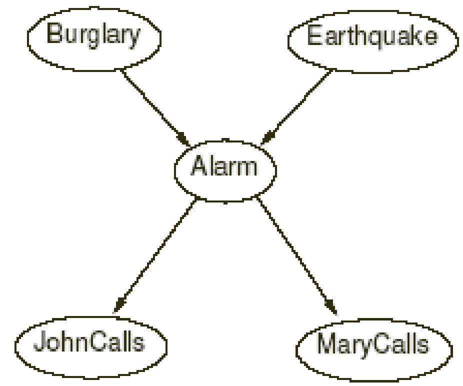
$$\cancel{P(+m \mid +a, -b, -e)} \xrightarrow{M \perp \{B,E\} \mid A} P(+m \mid +a)$$

$$\cancel{P(+a \mid -b, -e)} \longrightarrow P(+a \mid -b, -e)$$

$$\cancel{P(-b \mid -e)} \xrightarrow{B \perp E} P(-b)$$

$$\cancel{P(-e)} \longrightarrow P(-e)$$

Joint Distribution



Node is INDEPENDENT of non-descendants, given assignments to its parents

P(+j, +m, +a, -b, -e)

= P(+j | +a)

P(+m | +a)

P(+a| -b, -e)

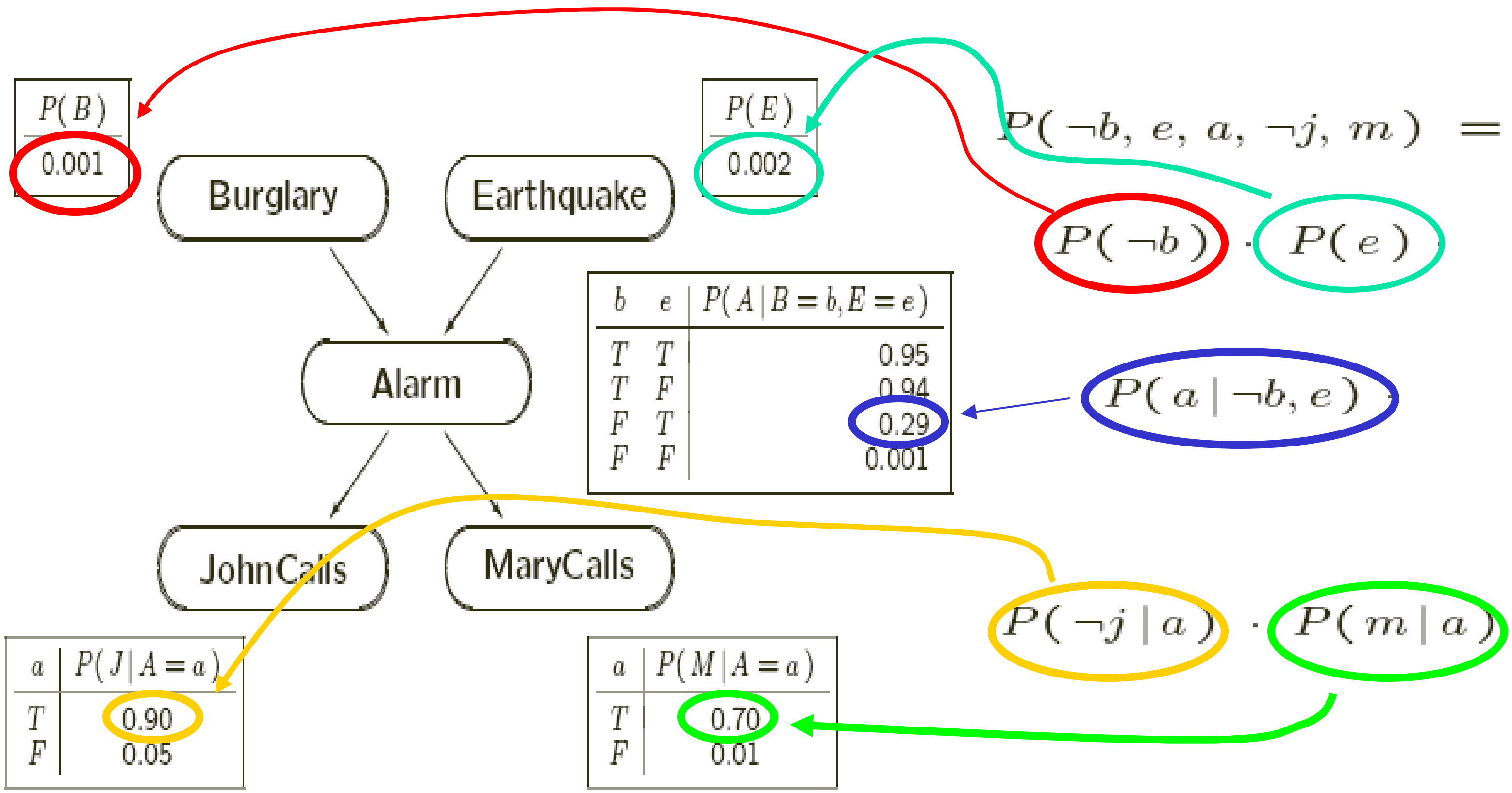
P(-b)

P(-e)

Recovering Joint

$$
\begin{aligned}
P(\neg b, e, a, \neg j, m) \quad = & \\
& P(\neg b)\ P(e \mid \neg b)\ P(a \mid e, \neg b)\ P(\neg j \mid a, e, \neg b)\ P(m \mid \neg j, a, e, \neg b) \\
& P(\neg b)\ P(e)\ \quad P(a \mid e, \neg b)\ P(\neg j \mid a)\ \quad P(m \mid a) \\
& 0.99 \times 0.02 \times \quad 0.29 \times \quad 0.1 \times \quad 0.70
\end{aligned}
$$

$\boxed{\text{Node independent of predecessors, given parents}}$



$P(B)$
0.001

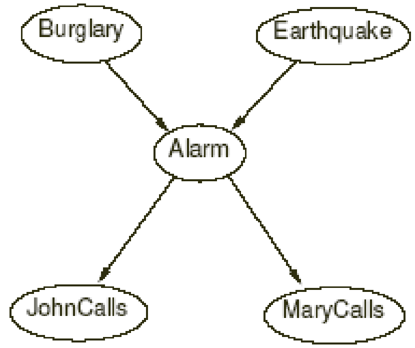
$P(E)$
0.002

b	e	$P(A \mid B=b, E=e)$
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

a	$P(J \mid A=a)$
T	0.90
F	0.05

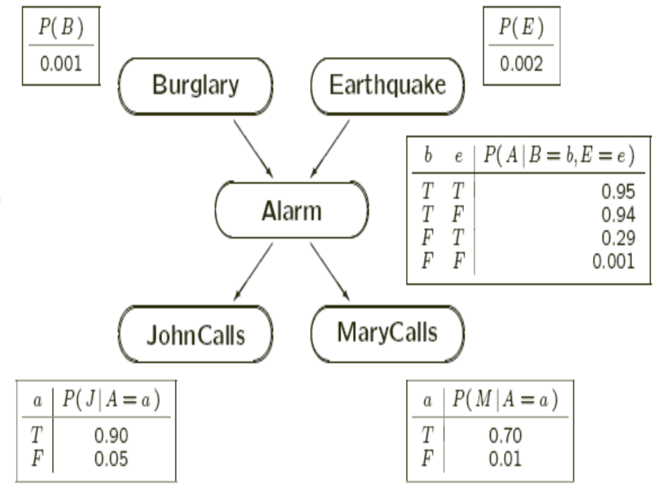
a	$P(M \mid A=a)$
T	0.70
F	0.01

Meaning of Belief Net



- A BN represents
  - joint distribution
  - condition independence statements
- P(+j, +m, +a, -b, -e)
  = P(-b) P(-e) P(+a|-b, -e) P(+j | +a) P(+m |+a)
  = $0.999 \times 0.998 \times 0.001 \times 0.90 \times 0.70 = 0.00062$
- In gen'l, $P(X_1, X_2, \ldots, X_m) = \prod_i P(X_i \mid X_{i+1}, \ldots, X_m)$
- Independence means
  $P(X_i \mid X_{i+1}, \ldots, X_m) = P(X_i \mid \text{Parents}(X_i))$
  Node independent of predecessors, given parents
- So... $P(X_1, X_2, \ldots, X_m) = \prod_i P(X_i \mid \text{Parents}(X_i))$

Comments



- BN used 10 entries

  ... can recover full joint (2^5 entries)
  - Given structure,
    other $2^5 - 10$ entries are REDUNDANT

$\Rightarrow$ Can compute

P(+burglary | +johnCalls, -maryCalls) :

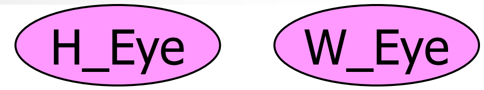
Get joint, then marginalize, conditionalize, ...

$\exists$ better ways. . .

- Note: Given structure, ANY CPT is consistent.

  $\nexists$ redundancies in BN. . .

"V"-Connections

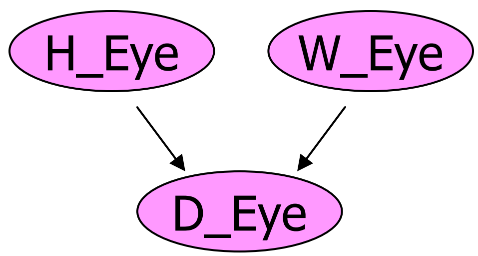


- What color are my wife's eyes?
- Would it help to know MY eye color?

  NO! H_Eye and W_Eye are independent!

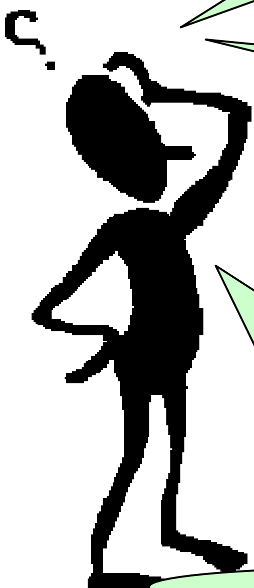
- We have a DAUGHTER, who has BROWN eyes

  Now do you want to know my eye_color?



h	w	P(D= bl | h , w)
bl	bl	1.0
bl	br	0.5
br	bl	0.5
br	br	0.25

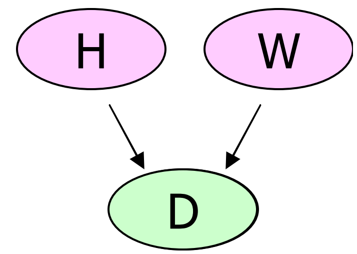
- H_Eye and W_Eye became dependent!



What color is W?

Prior is $P(W = br) = 0.8$?

But I know H!
Should I tell you?



Don't bother; it doesn't matter

$P(W = br \mid H = bl) = 0.8$

$P(W = br \mid H = br) = 0.8$

I also know $D = br$. Now do you care?

Yes, yes!!! Tell me H!

$P(W = br \mid H = bl, D = br) = 0.50$

$P(W = br \mid H = br, D = br) = 0.22$

d-separation Conditions

$\neg(X \perp Y)$

$X \rightarrow Z \rightarrow Y$

Earthquake $\rightarrow$ Alarm $\rightarrow$ JohnCalls

$\neg(X \perp Y)$

$X \leftarrow Z \rightarrow Y$

MaryCalls $\leftarrow$ Alarm $\rightarrow$ JohnCalls

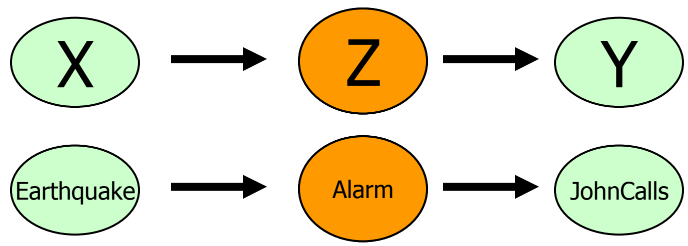
$X \perp Y$

$X \rightarrow Z \leftarrow Y$

Earthquake $\rightarrow$ Alarm $\leftarrow$ Burglary

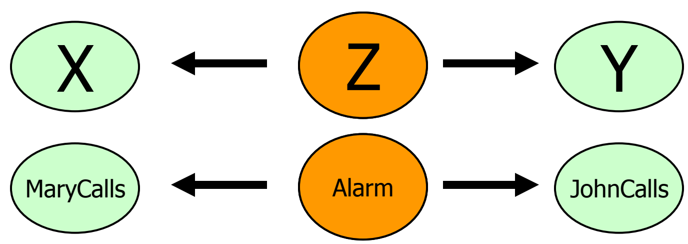
d-separation Conditions

$\neg(X \perp Y)$



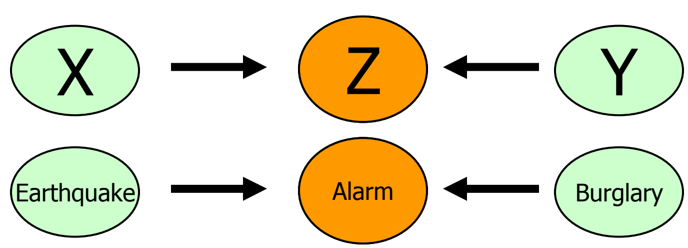
$X \perp Y \mid Z$

$\neg(X \perp Y)$



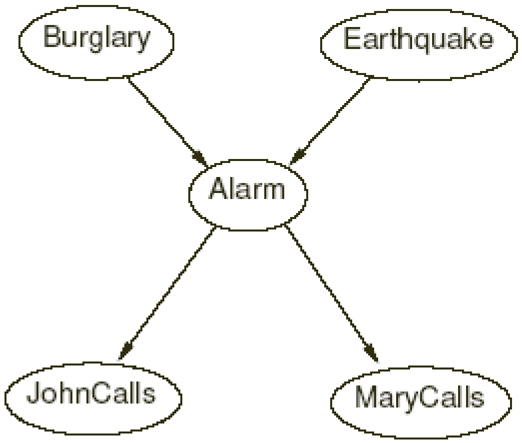
$X \perp Y \mid Z$

$X \perp Y$



$\neg(X \perp Y \mid Z)$

d-separation



- Burglary and JohnCalls are conditionally independent given Alarm
- JohnCalls and MaryCalls are conditionally independent given Alarm
- Burglary and Earthquake are independent given no other information
- But ...
  - Burglary and Earthquake are dependent given Alarm
  - Ie, Earthquake may "explain away" Alarm … decreasing prob of Burglary

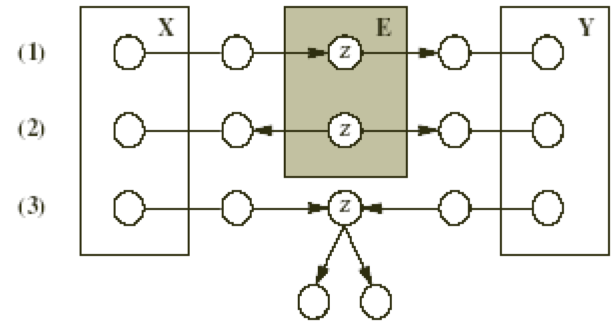
Conditional Independence

- Node X is independent of its non-descendants given assignment to immediate parents parents(X)

- **General** question: "$X \perp Y \mid \mathbf{E}$"
  - Are nodes **X** independent of nodes **Y**, given assignments to (evidence) nodes **E**?

- **Answer**: If every undirected path from X to Y is d-separated by **E**, then $X \perp Y \mid \mathbf{E}$

- *d-separated* if every path from X to Y is blocked by **E**
  . . . if ∃ node Z on path s.t.
  1. $Z \in \mathbf{E}$, and Z has 1 out-link (on path)
  2. $Z \in \mathbf{E}$, and Z has 2 out-link, *or*
  3. Z has 2 in-links, $Z \notin \mathbf{E}$, no child of Z in **E**

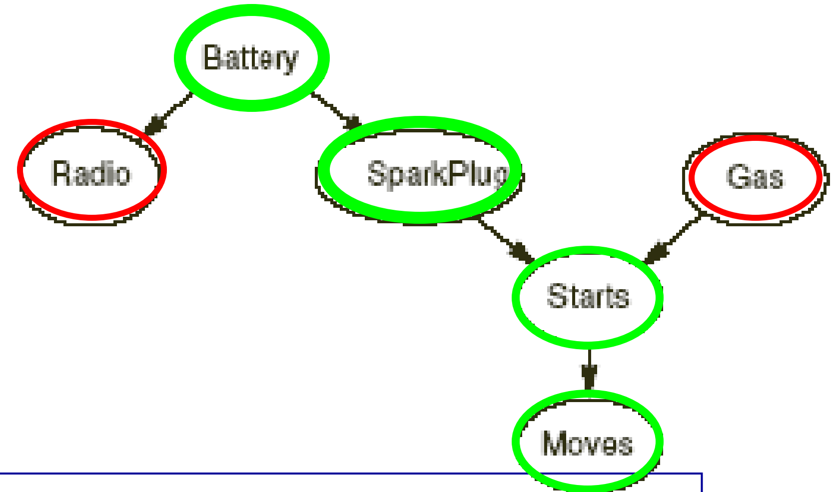


Conditional Dependence

- Node X is independent of its non-descendants given assignment to immediate parents parents(X)
- **General** question: "$\neg(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{E})$ "
  - Are nodes **X** dependent of nodes **Y**, given assignments to (evidence) nodes **E**?
- **Answer**: $\neg(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{E})$ if
  any undirected path from **X** to **Y**
  is *active* given **E**
- iff...
  1. whenever node Z on path has 2 in-links, $Z \in \mathbf{E}$ or some child of Z in **E**
  2. no other node Z is in **E**

Example of *Active Path*

"flow" if
any path from X to Y is active wrt **E**



Any flow from *Radio* to *Gas* given ...

1. **E** = {} ?

   No: $P(R \mid G) = P(R)$

   Starts $\notin$ **E**, and Starts has 2 in-links

2. **E** = Starts ?

   YES!! $P(R \mid G, S) \neq P(R \mid S)$

   Starts $\in$ **E**, and Starts has 2 in-links

3. **E** = Moves ?

   YES!! $P(R \mid G, M) \neq P(R \mid M)$

   Moves $\in$ **E**, Moves child-of Starts, and Starts has 2 in-links (on path)

4. **E** = SparkPlug ?

   NO: $P(R \mid G, Sp) = P(R \mid Sp)$

   SparkPlug $\in$ **E**, and SparkPlug has 1 out-link

5. **E** = Battery ?

   NO: $P(R \mid G, B) = P(R \mid B)$

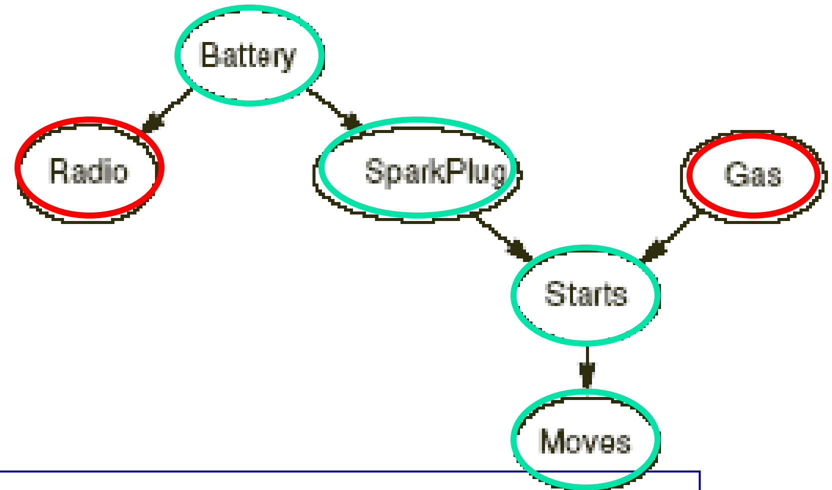
   Battery $\in$ **E**, and Battery has 2 out-links

If car does not start

If car does not MOVE,
expect radio to NOT work.
Unless you see it is out of gas!

Example of *d*-separation

d-separated if
every path from X to Y is blocked by **E**



Is Radio *d*-separated from Gas given . . .

1. **E** = {} ?
   YES: P(R | G) = P(R)
   Starts ∉ **E**, and Starts has 2 in-links
2. **E** = Starts ?
   NO!! P(R | G, S) ≠ P(R| S)
   Starts ∈ **E**, and Starts has 2 in-links
3. **E** = Moves ?
   NO!! P(R | G, M) ≠ P(R| M)
   Moves ∈ **E**, Moves child-of Starts, and Starts has 2 in-links (on path)
4. **E** = SparkPlug ?
   YES: P(R | G, Sp) = P(R| Sp)
   SparkPlug ∈ **E**, and SparkPlug has 1 out-link
5. **E** = Battery ?
   YES: P(R | G, B) = P(R| B)
   Battery ∈ **E**, and Battery has 2 out-links

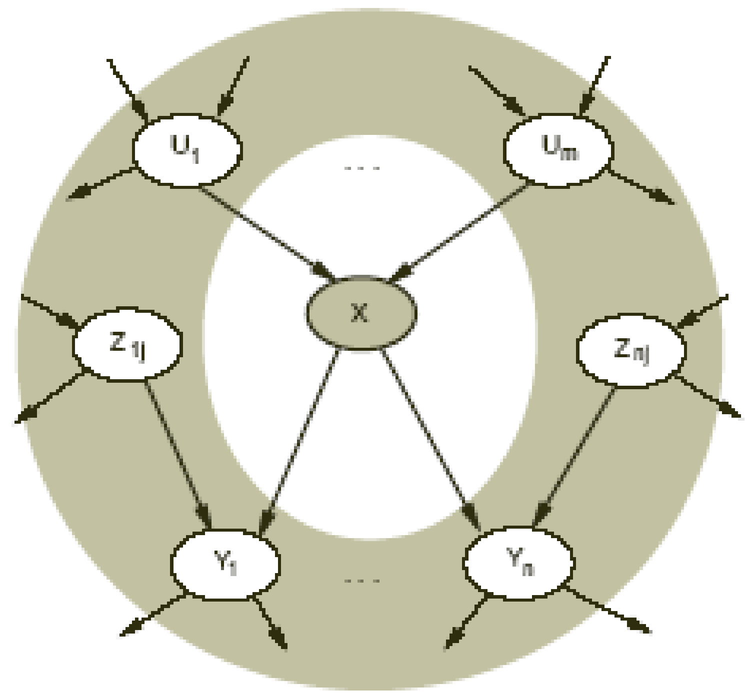
If car does not start

If car does not MOVE,
expect radio to NOT work.
Unless you see it is out of gas!

Markov Blanket

Each node is
conditionally independent of all others
given its *Markov blanket:*

- parents
- children
- children's parents



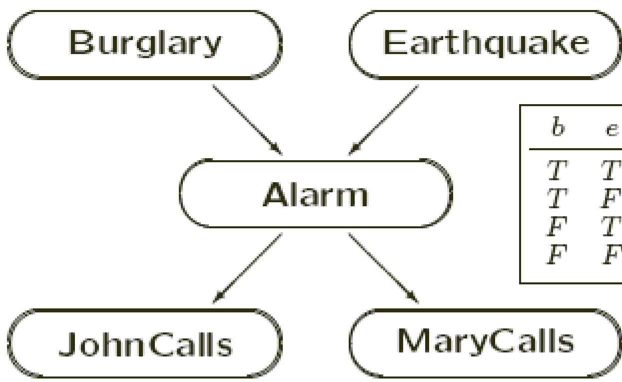
Example Bayesian Net

Directed Acyclic Graph:

$$\mathcal{BN} = \left\{ \begin{array}{lll} \mathcal{N} & \text{Nodes} & \equiv \text{Variables} \\ \mathcal{A} & \text{Arcs} & \equiv \text{Dependencies} \\ \mathcal{C} & \text{CPTables} & \equiv \text{“weights”} \end{array} \right\}$$

$P(B)$
0.001

$P(E)$
0.002



b	e	$P(A \mid B=b, E=e)$
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

- Discrete variables
- Explicit table

a	$P(J \mid A=a)$
T	0.90
F	0.05

a	$P(M \mid A=a)$
T	0.70
F	0.01

- **Nodes**: one for each random variable
- **Arcs**: one for each direct influence between two r.v.s
- **CPT**: each node stores a conditional probability table

  P(Node | Parents(Node))

  to quantify effects of “parents” on child

Skip

Simple forms of CPTable

- In gen'l: CPTable is function mapping
  values of parents to *distribution over child*

$$f: \left[\prod_{U \in Parents(X)} Dom(U) \right] \times Dom(X) \mapsto [0,1]$$

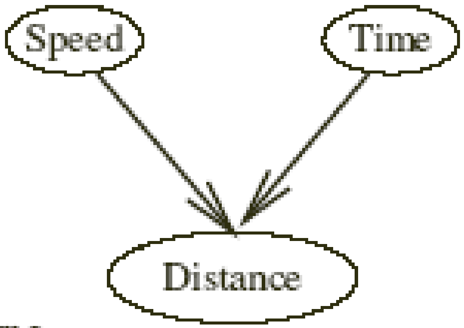
(Actually, $f': \prod_{U \in Parents(X)} Dom(U) \mapsto$ dist over X)

Cold	Flu	Malaria	$P($Fever$\mid$C,F,M$)$	$P(\neg$Fever$\mid$C,F,M$)$
F	F	F	0.0	1.0
F	F	T	0.9	0.1
F	T	F	0.8	0.2
F	T	T	0.98	0.02
T	F	F	0.4	0.6
T	F	T	0.94	0.06
T	T	F	0.88	0.12
T	T	T	0.988	0.012

- Standard: Include $\prod_{U \in Parents(X)} |Dom(U)|$ rows, each with $|Dom(X)| - 1$ entries
- But... can be structure within CPTable: Deterministic, Noisy-Or, (Decision Tree), ...

Deterministic Node

- Given value of parent(s), specify unique value for child (logical, functional)



$$P(\texttt{Distance} \mid \texttt{Rate},\ \texttt{Time}) = \begin{cases} 1.0 & \text{if } \texttt{Distance} = \texttt{Rate} \cdot \texttt{Time} \\ 0.0 & \text{otherwise} \end{cases}$$

As if each row has just one 1, rest 0s:

Rate	Time	P(Dist=0 | R,T)	P(Dist=1 | R,T)	P(Dist=2 | R,T)
0	1	**1.0**	0.0	0.0
1	0	**1.0**	0.0	0.0
1	1	1.0	**1.0**	0.0
1	2	0.0	0.0	**1.0**
2	1	0.0	0.0	**1.0**
⋮		⋮		

Noisy-OR CPTable

- Each cause is independent of the others
- All possible causes are listed

Want: No Fever if none of Cold, Flu or Malaria

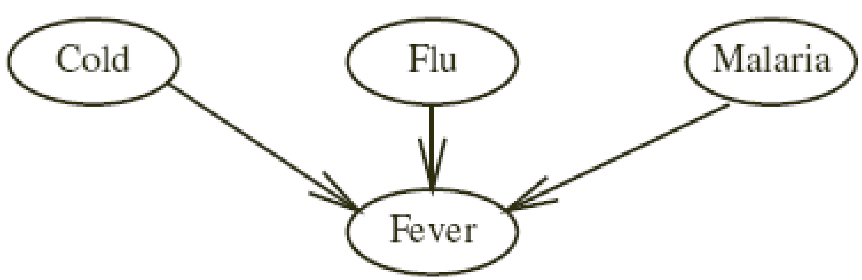
$P(\neg Fev \mid \neg Col, \neg Flu, \neg Mal) = 1.0$

\+ Whatever inhibits cold from causing fever

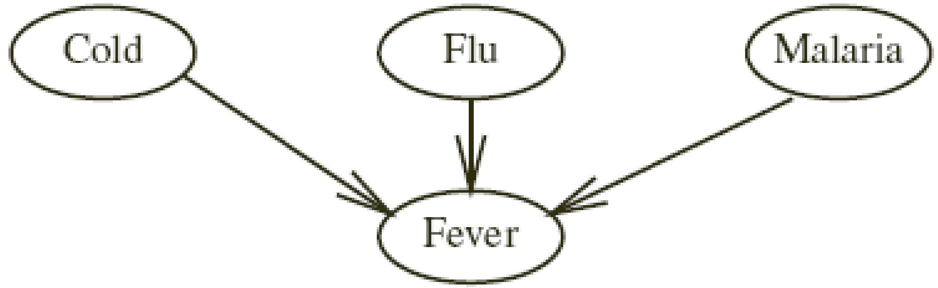
is independent of

whatever inhibits flu from causing fever

$P(\neg Fev \mid Cold, Flu) \approx P(\neg Fev \mid Cold)\, P(\neg Fev \mid Flu)$



Noisy-OR "CPTable" (2)



- $P(\texttt{Fev} \mid \neg\texttt{Col}, \neg\texttt{Flu}, \neg\texttt{Mal}) = 0$

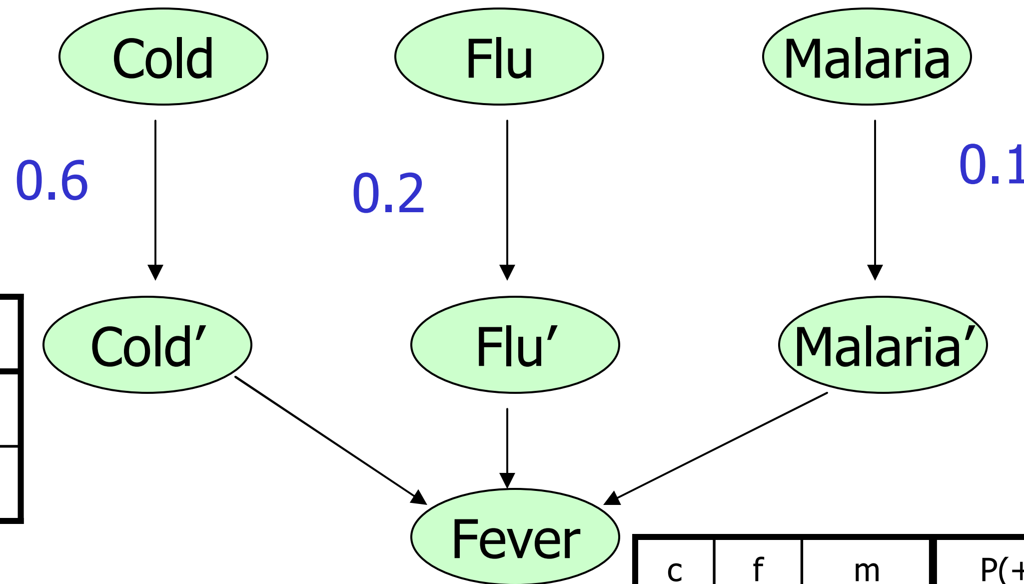
$$P(\neg\texttt{Fev} \mid \texttt{Col}) \approx q_{col} = 0.6$$
$$P(\neg\texttt{Fev} \mid \texttt{Flu}) \approx q_{flu} = 0.2$$
$$P(\neg\texttt{Fev} \mid \texttt{Mal}) \approx q_{mal} = 0.1$$

- Independent inhibiters:
  $P(\neg\texttt{Fev} \mid \texttt{Col}, \texttt{Flu}) \approx P(\neg\texttt{Fev} \mid \texttt{Col}) \times P(\neg\texttt{Fev} \mid \texttt{Flu})$

$$P(\neg\texttt{Fever} \mid \pm_i d_i) = \prod_{i:+d_i} q_i$$

Cold	Flu	Malaria	$P(\neg\texttt{Fever} \mid \texttt{c,f,m})$	$P(\texttt{Fever} \mid \texttt{c,f,m})$
F	F	F	1.0	0.0
F	F	T	0.1	0.9
F	T	F	0.2	0.8
F	T	T	$0.02 = 0.2 \times 0.1$	0.98
T	F	F	0.6	0.4
T	F	T	$0.06 = 0.6 \times 0.1$	0.94
T	T	F	$0.12 = 0.6 \times 0.2$	0.88
T	T	T	$0.012 = 0.6 \times 0.2 \times 0.1$	0.988

Noisy-Or … expanded



c	P(+cold' | c)	P(-cold' | c)
+	$1\text{-}q_c = 0.4$	$q_c = 0.6$
–	0.0	1.0

c	f	m	P(+Fever|c, f, m)
+	+	+	1.0
+	+	–	1.0
+	–	+	1.0
+	–	–	1.0
+	+	+	1.0
–	+	–	1.0
–	–	+	1.0
–	–	–	0.0

Noisy-Or (Gen'l)

- Fever if Cold, Flu or Malaria

$$\text{Want} \begin{cases} P(\texttt{Fev} \mid \neg\texttt{Col}, \neg\texttt{Flu}, \neg\texttt{Mal}) = 0 \\ P(\neg\texttt{Fev} \mid \texttt{Col}) \approx q_{col} = 0.6 \\ P(\neg\texttt{Fev} \mid \texttt{Flu}) \approx q_{flu} = 0.2 \\ P(\neg\texttt{Fev} \mid \texttt{Mal}) \approx q_{mal} = 0.1 \end{cases}$$

("noise" parameters)

CPCS Network:

- Modeling disease/symptom for internal medicine
- Using Noisy-Or & Noisy-Max
- 448 nodes, 906 links
- Required 8,254 values (not 13,931,430) !

Assumes: – each cause has ... effect

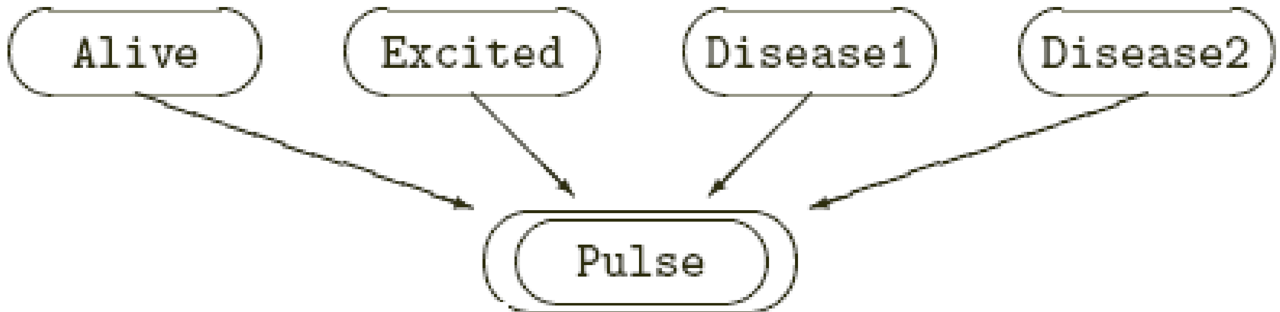
– all causes listed

(Leak node, to handle ALL ... ...)

– inhibiting factors independent

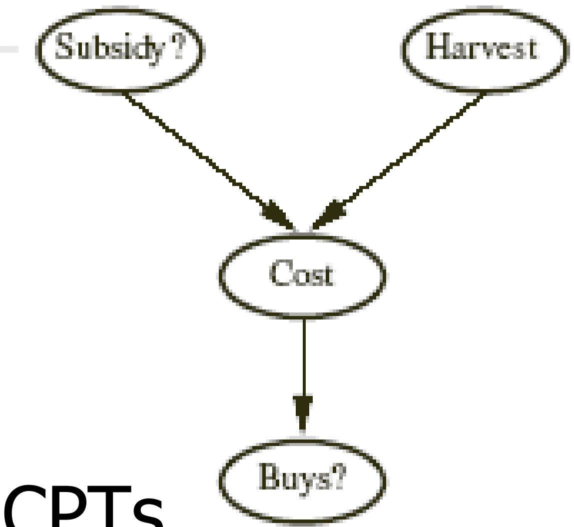
Note: Only k parameters, not 2^k

DecisionTree CPTable



A	E	D1	D2	χ s.t. $P(\texttt{Pulse}=\chi \mid \texttt{A,E,D1,D2}) = 1.0$
Y	Y	Y	Y	vhigh
Y	Y	Y	N	vhigh
Y	Y	N	Y	vhigh
Y	Y	N	N	vhigh
Y	N	Y	Y	high
Y	N	Y	N	med
Y	N	N	Y	med
Y	N	N	N	ok
N	Y	Y	Y	none
N	Y	Y	N	none
N	Y	N	Y	none
N	Y	N	N	none
N	N	Y	Y	none
N	N	Y	N	none
N	N	N	Y	none
N	N	N	N	none

Hybrid (discrete+continuous) Networks



- **Discrete**: Subsidy?, Buys?

  Continuous: Harvest, Cost

Option 1: Discretization

but possibly large errors, large CPTs

Option 2: Finitely parameterized canonical families

Problematic cases to consider. . .

- Continuous variable, discrete+continuous parents

  Cost

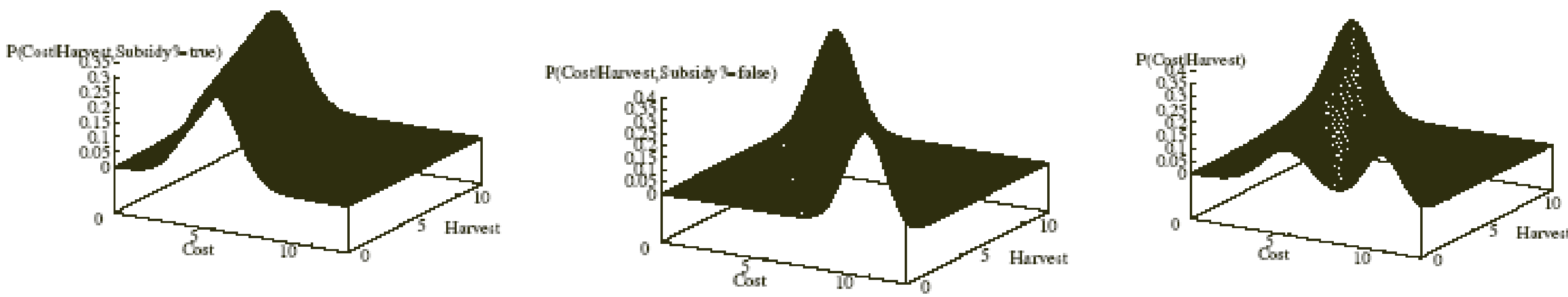
- Discrete variable, continuous parents

  Buys?

Skip

If everything is Gaussian...

- All nodes continuous w/ LG dist'ns
  ⇒ full joint is a multivariate Gaussian

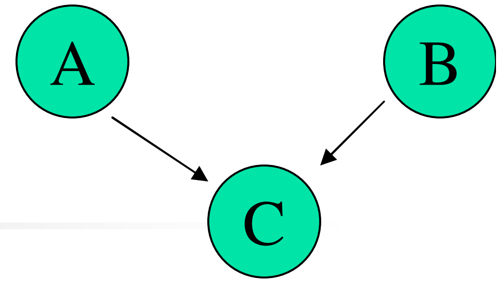


- Discrete+continuous LG network

⇒ conditional Gaussian network

multivariate Gaussian over all continuous variables
for each combination of discrete variable values

Linear Gaussian Model

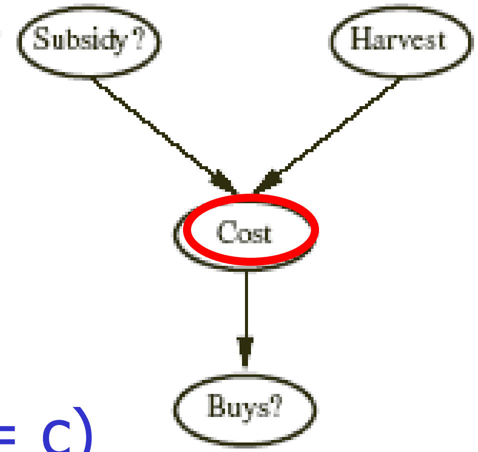


- $P(x_i \mid pa_i) \sim \mathbb{N}(x_i \mid b_i + \sum_{j \in pa\_i} w_{ij}\, x_j,\ \ v_i)$
- So…
  - $P(x_A) \sim \mathcal{N}(x_A \mid b_A,\ \ v_A)$
  - $P(x_B) \sim \mathcal{N}(x_B \mid b_B,\ \ v_B)$
  - $P(x_C \mid x_A, x_B) \sim \mathcal{N}(x_C \mid b_C + w_{AC}\, x_A + w_{BC}\, x_B,\ \ v_C)$

  … eg, $\mathbb{N}(x_C \mid 2.9 + 1.3\, x_A + -21\, x_B,\ \ 0.5)$
- $\ln p(\mathbf{x}) = \sum_i \ln p(x_i \mid pa_i) =$

$$-\sum_i \frac{1}{2v_i}\left(x_i - \sum_{j \in pa_i} w_{ij} x_i - b_i\right)^2 + const.$$

Continuous Child Variables



- For each "continuous" child E,
  - with continuous parents C
  - with discrete parents D
- Need conditional density function

  P(E = e | C = c, D = d) = $P_{D=d}$(E = e | C = c)

  for each assignment to discrete parents D=d
- Common: linear Gaussian model

  f(Harvest, Subsidy?) = "dist over Cost"

$$P(\texttt{Cost} = \texttt{c} \mid \texttt{Harvest}=\texttt{h},\ \texttt{Subsidy?}=\texttt{true})$$
$$= \mathcal{N}[a_t h + b_t,\ \sigma_t](c)$$
$$= \frac{1}{\sigma_t\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{c-(a_t h + b_t)}{\sigma_t}\right)^2\right)$$
$$P(\texttt{Cost} = \texttt{c} \mid \texttt{Harvest}=\texttt{h},\ \texttt{Subsidy?}=\texttt{false})$$
$$= \mathcal{N}[a_f h + b_f,\ \sigma_f](c)$$

Need parameters:

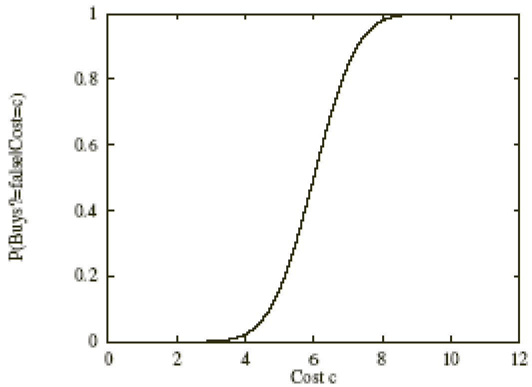
σ_t a_t b_t

σ_f a_f b_f

Discrete variable w/ Continuous Parents

- Probability of Buys? given Cost

  $\approx^?$ "soft" threshold:



- Probit distribution uses integral of Gaussian:

$$\Phi(x) \quad = \quad \int_{-\infty}^{x} \mathcal{N}[0,1](x)\, dx$$

$$P(\texttt{Buys?}=\texttt{true}\,|\,\texttt{Cost}=c) \quad = \quad \Phi\left(\frac{\mu-c}{\sigma}\right)$$

$\approx$ hard threshold, whose location is subject to noise

Outline

- Motivation
- What is a Belief Net?
  - Example
  - Inference
  - Semantics
  - Relation to other Models
    - Rules, Neural Nets, Markov Nets, Clusters
- Learning a Belief Net

- My Research

Belief Nets vs Rules

- Both have "*Locality*"
  Specific clusters (rules / connected nodes)
- Often *same nodes* (rep'ning Propositions) but

BN:	Cause	$\Rightarrow$	Effect	
	"Hep	$\Rightarrow$	Jaundice"	$P(J \mid H)$
Rule:	Effect	$\Rightarrow$	Cause	
	"Jaundice	$\Rightarrow$	Hep"	

WHY?: Easier for people to reason ***CAUSALLY***
even if use is ***DIAGNOSTIC***

- BN *provide OPTIMAL way to deal with*
  + ***Uncertainty***
  + ***Vagueness*** ***(var not given, or only dist)***
  + ***Error***

...Signals meeting Symbols ...

- BN *permits* different "direction"'s of inference

Belief Nets vs Neural Nets

- Both have "*graph structure*" but

BN:	Nodes have SEMANTICs Combination Rules: Sound Probability
NN:	Nodes: arbitrary Combination Rules: Arbitrary

- So harder to
  - *Initialize NN*
  - *Explain NN*

  (But perhaps easier to learn NN from examples only?)
- BNs can deal with
  - *Partial Information*
  - *Different "direction"s of inference*

Belief Nets vs Markov Nets

- Each uses "*graph structure*"

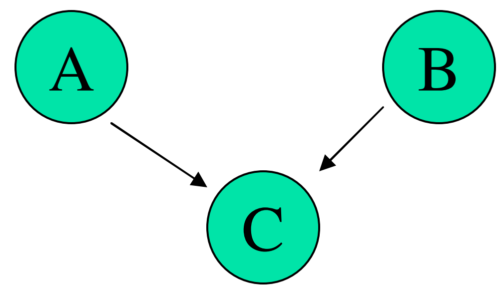
  to FACTOR a distribution

  … explicitly specify dependencies, implicitly independencies…

- but subtle differences…
  - BNs capture "causality", "hierarchies"
  - MNs capture "temporality"

Technical: BNs use DIRECTRED arcs

$\Rightarrow$ allow "induced dependencies"

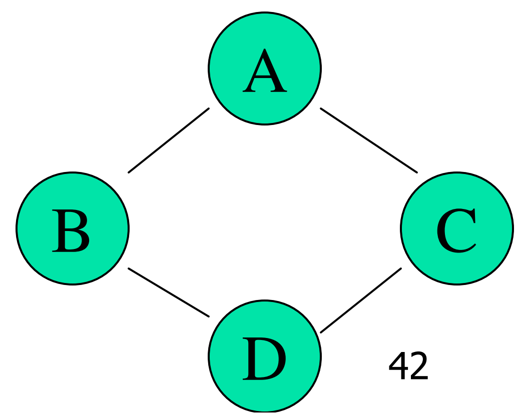


$I\,(A, \{\}, B)$ "A independent of B, given $\{\}$"

$\neg\, I\,(A, C, B)$ "A dependent on B, given C"

MNs use UNDIRECTED arcs

$\Rightarrow$ allow other independencies

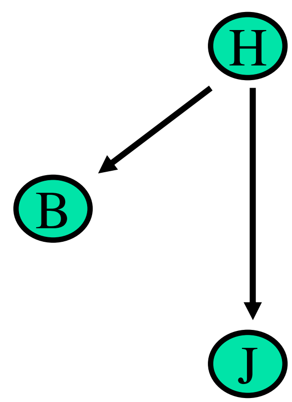


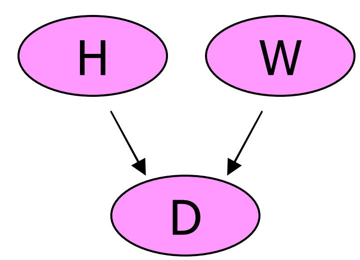
$I(A, BC, D)$ A independent of D, given B, C

$I(B, AD, C)$ B independent of C, given A, D

Belief Nets vs Clusters

- Both "structure" the variables
  - Cluster: Put *similar* variables in same cluster
  - BN: Put *related* variables adjacent
- Cluster uses "first order" relationships
  - Put A and B together if A *directly correlated with* B
- BN can have higher order relationships, esp. independencies





2^{nd} Order Statistics?

- Spse
  - ½ of kidney *donors* are Male (½ female)
  - ½ of kidney *recipients* are Male (½ female)
  - Transplant is SUCCCESSFUL iff
    Donor and Recipient are SAME gender (M/M or F/F)
- Here:
  - P(Success | Donor=m) = ½ = P(Success | Donor=f)
    ⇒ Success is independent of Donor Gender
  - P(Success | Recip=m) = ½ = P(Success | Recip=f)
    ⇒ Success is independent of Recipient Gender
- However:
  - P(Success | Donor=m, Recip=f) = 0
    P(Success | Donor=m, Recip=m) = 1
  - So Success is dependent on
    Recipient Gender and Donor Gender

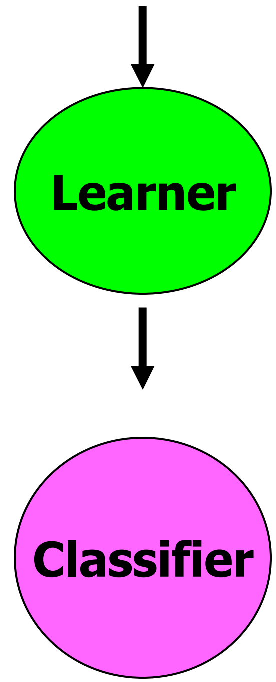
Outline

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

- My Research

Learning is … Training a Classifier

Temp.	Press.	Sore Throat	…	Colour	diseaseX
35	95	Y	…	Pale	No
22	110	N	…	Clear	Yes
:	:			:	:
10	87	N	…	Pale	No

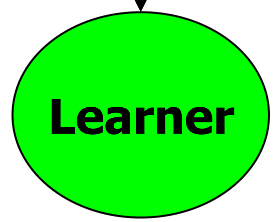


Temp	Press.	Sore-Throat	…	Color
32	90	N	…	Pale

diseaseX
No

Learning is … Training a Model

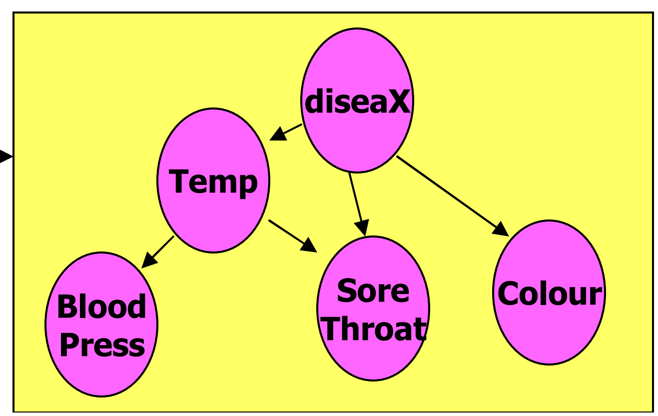
Temp.	Blood Press.	Sore Throat	…	Colour	diseaseX
35	95	Y	…	Pale	No
22	110	N	…	Clear	Yes
:	:			:	:
10	87	N	…	Pale	No



Then conditionalize, marginalize to answer *any question*:

P(+d | temp=30, BP= 100, …)

Temp	Blood Press.	Sore-Throat	…	Color	diseaseX
32	90	N	…	Pale	No



J	H	B	P(j,b,h)
0	0	0	0.03395
0	0	1	0.0095
0	1	0	0.0003
0	1	1	0.1805
1	0	0	0.01455
1	0	1	0.038
1	1	0	0.00045
1	1	1	0.722

Why Learn?
Why not just "program it in"?

Appropriate Model …

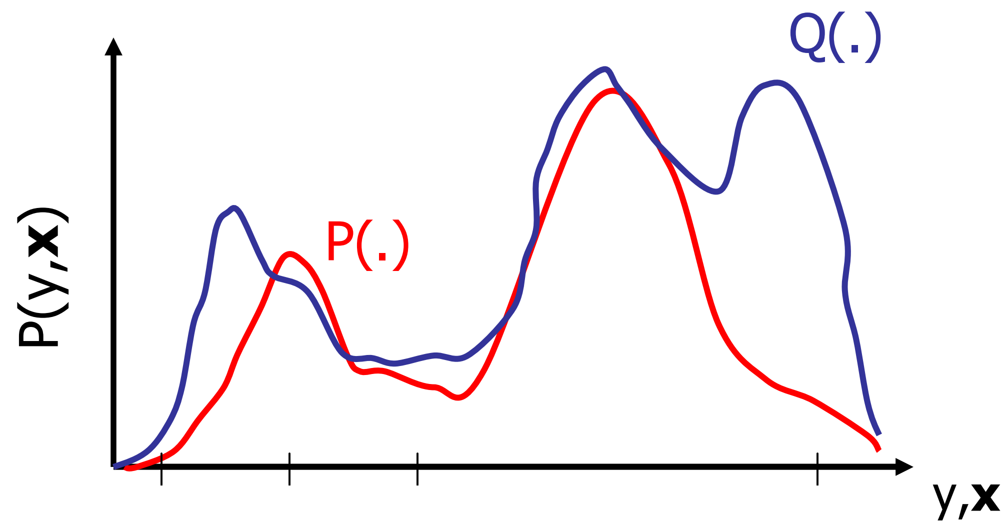
- … is not known
  Medical diagnosis… Credit risk… Control plant…
- … is too hard to "engineer"
  Drive a car… Recognize speech…
- … changes over time
  Plant evolves…
- … user specific
  Adaptive user interface…

Why Learn Bayes Nets?

- Goal#1: Build a classifier
  - What is P(Cancer = + | HA = +, Fev = –, …) ?
  - Is P(Cancer = + | …) > P(Cancer = – | …) ?
- Goal#2: Build a SET of classifiers
  - What is P(Cancer = + | HA = +, Fev = –, …) ?
  - What is P(Meningitis = – | HA = +, Cold = 3, …) ?
  - What is P(HospStay = 3 | Smoke = 0.1, BNose = –1, …) ?
- Goal#3: Build a model of the world!
  - . . . all interrelations between all subsets of variables
  - Reveal (in)dependencies, connections, …
  - "Density Estimation"
  - Note: A completely accurate model will produce correct answers to EVERY P(X | Y) query

Generative vs Discriminative

- Generative Learning:
  - Given (sample of) distribution, P(y,**x**)
  - Seek model Q(y,**x**) that matches P(y,**x**)



- Discriminative Learning:
  - Given (sample of) distribution, P(y,**x**)
  - Seek model Q(y | **x**) that matches P(y | **x**)

S	A	⋯	G	C_P	C_Q
y	y	⋯	m	1	1
n	o	⋯	f	1	0
y	o	⋯	f	0	0
⋮	⋮		⋮	⋮	⋮

KL-Divergence … ≈ MaxLikelihood

- Seek the BN that minimizes KL-divergence

$$\boxed{KL(\ D;\ BN\ ) = \sum_x P_D(x) \ln \frac{P_D(x)}{P_{BN}(x)}}$$

- KL-divergence …
  - always ≥ 0
  - =0 iff distr's "identical"
  - not symmetric
- but… distrib'n $\mathcal{D}$ not known;
  Only have instances
  $S = \{d_r\}$
  drawn iid from $\mathcal{D}$

$$
\begin{aligned}
\bullet\ BN^* &= \underset{BN}{\text{argmin}}\ KL(\mathcal{D};\ BN\ ) \\
&= \underset{BN}{\text{argmax}} \sum_x P_{\mathcal{D}}(x) \ln P_{BN}(x) \\
&\approx \underset{BN}{\text{argmax}} \frac{1}{|S|} \sum_{d \in S} \ln P_{BN}(d) \\
&= \underset{BN}{\text{argmax}} \prod_{d \in D} P_{BN}(d) \quad = \quad \underset{BN}{\text{argmax}}\ P_{BN}(S)
\end{aligned}
$$

as $\sum_x P_D(x) \ln P_D(x)$ is independent of BN

as S drawn from D

Best Distribution

- If goal is
  BN that approximates $\mathcal{D}$:

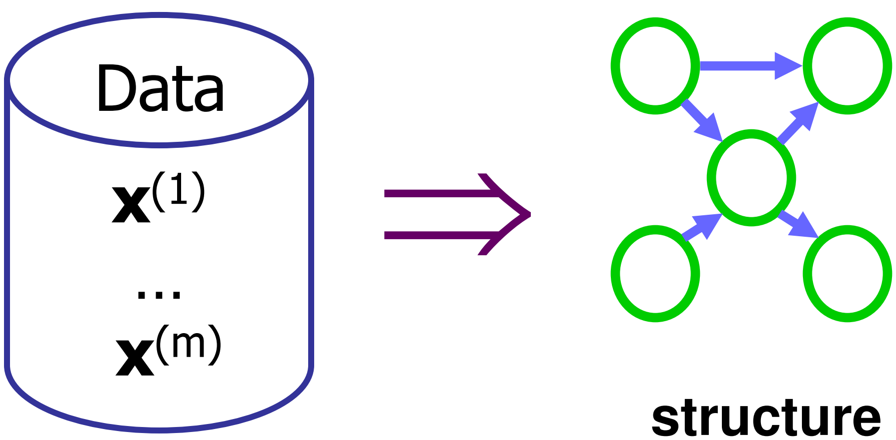
  Find BN^* that maximizes likelihood of data S

$$\boxed{\arg\min_{BN} KL(\ D;\ BN\ ) \ \approx\ \arg\max_{BN} P_{BN}(S)}$$

- Approaches:
  - Frequentist: *Maximize Likelihood*
    - to address overfitting: BDe, BIC, MDL, …
  - Bayesian: *Maximize a Posteriori*
  - …

Learning Bayes Nets

		Structure	
		Known	Unknown
Data	Complete	**Easy**	**NP-hard**
	Missing	**Hard … EM**	**Very hard!!**



structure

+

CPTs : $P(X_i \mid \mathbf{Pa}_{X_i})$

parameters

Typical (Benign) Assumptions

1. Variables are discrete
2. Each case $c_i \in \mathcal{S}$ is complete
3. Rows of CPtable are independent

   $\theta_A \perp \theta_B$

   $\theta_{B|+a} \perp \theta_{B|-a}$

$\theta_{+a},\ \theta_{-a}$

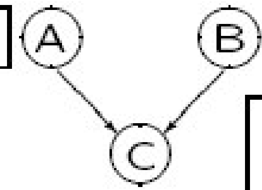
a	$P(B=+\mid A=a)$	$P(B=-\mid A=a)$
$+$	$\theta_{+b\|+a}$	$\theta_{-b\|+a}$
$-$	$\theta_{+b\|-a}$	$\theta_{-b\|-a}$

4. Prior $p(\Theta_\chi \mid \mathcal{G})$ is uniform
   - $\theta_{B|+a} \sim$ Beta(1,1)

- Later: relax Assumptions 1,2,4

Learning the CPTs

Structure: G =



$\theta_{+a} \quad \theta_{-a}$

$\theta_{+b} \quad \theta_{-b}$

$$\begin{array}{ll} \theta_{+c|+a,+b} & \theta_{-c|+a,+b} \\ \theta_{+c|-a,+b} & \theta_{-c|-a,+b} \\ \theta_{+c|+a,-b} & \theta_{-c|+a,-b} \\ \theta_{+c|-a,-b} & \theta_{-c|-a,-b} \end{array}$$

Sample S =

	A	B	C
d_1	1	0	1
d_2	0	1	0
d_3	0	0	1
d_4	1	0	1

- Given
  - Fixed structure $\mathcal{G}$
  - over discrete variables X_i
  - Complete instances $\mathcal{S}$
- $\widehat{\theta}$ = "empirical frequencies"
- Eg:
  - $\theta_{+a} = 2 / (2+2) = 0.5$
  - $\theta_{-b} = 3 / (3+1) = 0.75$
  - $\theta_{+c|+a,-b} = 2 / (2+0) = 1.0$

WHY????

One-Node Bayesian Net

- P(Heads) = θ, P(Tails) = 1-θ

C

P(C=h)	P(C=t)
θ	$1-\theta$

- Flips are i.i.d.:
  - Independent events
  - Identically distributed according to Binomial distribution
- Set $\mathcal{S}$ of α_H Heads and α_T Tails

$$\boxed{P(S \mid \theta) = \theta^{\alpha_H} (1-\theta)^{\alpha_T}}$$

Maximum Likelihood Estimation

- **Data:** Observed set $\mathcal{S}$ of
  α_H Heads and α_T Tails
- **Hypothesis Space:** Binomial distributions
- Learning θ is an optimization problem
  - What's the objective function?

- **MLE**: Choose θ that maximizes the probability of observed data:

$$\begin{aligned}\widehat{\theta} &= \arg\max_{\theta} \; P(\mathcal{S} \mid \theta) \\ &= \arg\max_{\theta} \; \ln P(\mathcal{S} \mid \theta)\end{aligned}$$

Simple "Learning" Algorithm

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta} \ \ln P(\mathcal{S} \mid \theta) \\ &= \arg\max_{\theta} \ \ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}\end{aligned}$$

- Set derivative to zero: $\boxed{\frac{d}{d\theta} \ \ln P(\mathcal{S} \mid \theta) = 0}$

$$\frac{\partial}{\partial\theta}\ln[\theta^h(1-\theta)^t] = \frac{\partial}{\partial\theta}[h\ln\theta + t\ln(1-\theta)] = \frac{h}{\theta} + \frac{-t}{(1-\theta)}$$

$$\frac{h}{\theta} + \frac{-t}{(1-\theta)} = 0 \Rightarrow \ \theta = \frac{h}{t+h}$$

So just average!!!

If 7 heads, 3 tails, set $\hat{\theta}$ = 0.7

Factoid wrt Belief Network

Recall that…

- For a COMPLETE instance, $\mathbf{x} = (x_1, \ldots, x_n)$

  $P(\mathbf{x})$ = product of CPtable values
  (one from each variable)

Probability of Complete Instance

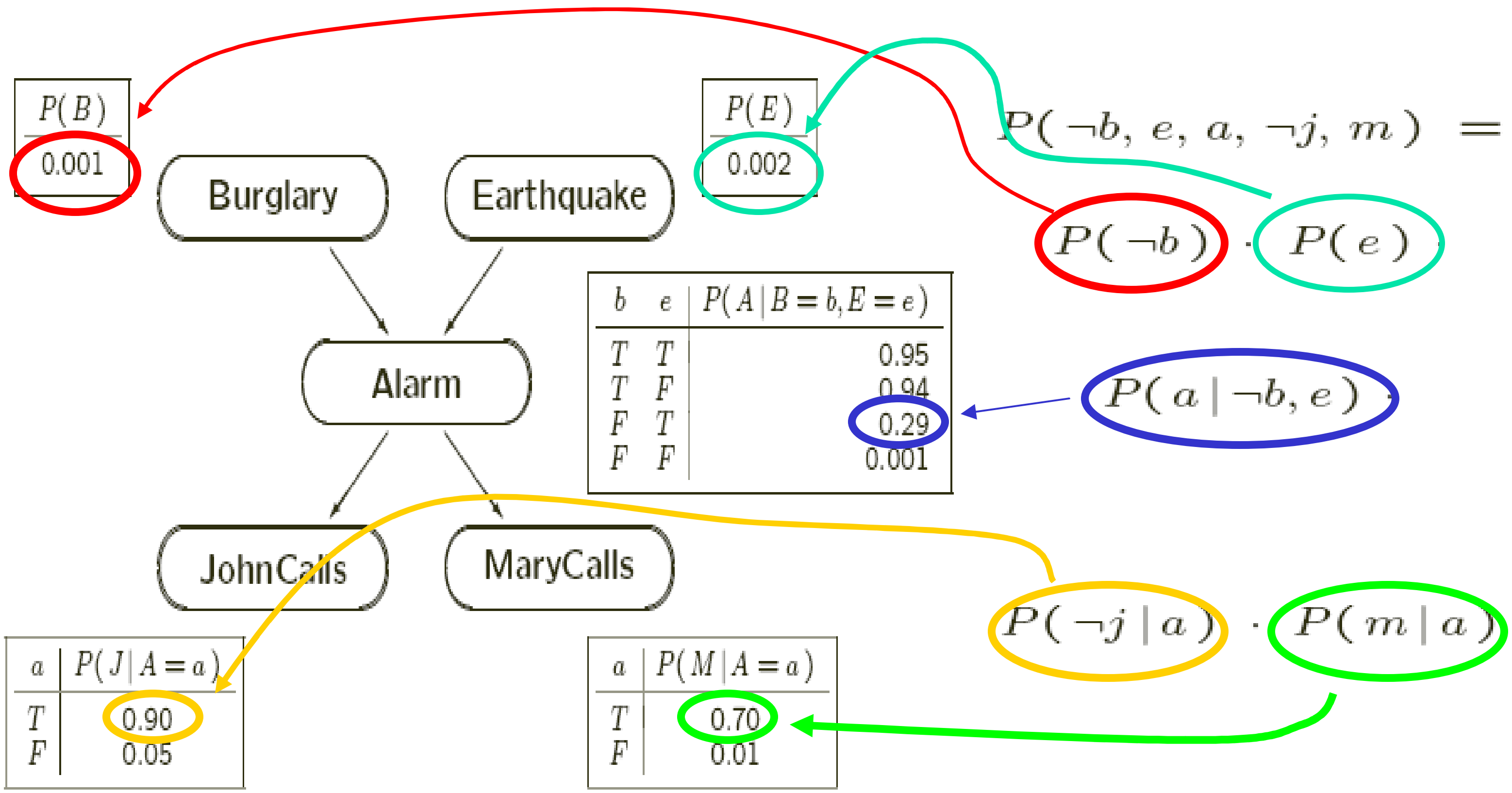
$$P(\neg b, e, a, \neg j, m) =$$

$$P(\neg b)\, P(e \mid \neg b)\, P(a \mid e, \neg b)\, P(\neg j \mid a, e, \neg b)\, P(m \mid \neg j, a, e, \neg b)$$

$$P(\neg b)\, P(e)\quad P(a \mid e, \neg b)\, P(\neg j \mid a)\quad P(m \mid a)$$

$$0.99 \times 0.02 \times 0.29 \times 0.1 \times 0.70$$

Node independent of predecessors, given parents



P(B)
0.001

P(E)
0.002

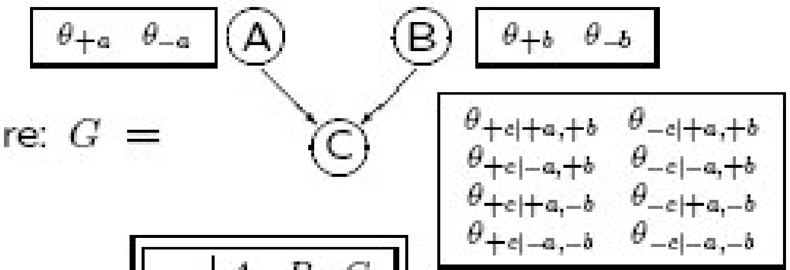
b	e	P(A | B = b, E = e)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

a	P(J | A = a)
T	0.90
F	0.05

a	P(M | A = a)
T	0.70
F	0.01

Likelihood of the Data (Frequentist)

Given: Structure: G =



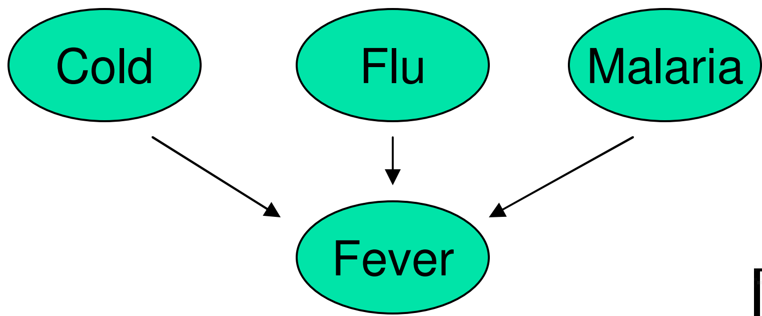
- $P(\mathcal{S} \mid \Theta) = \prod_r P(d_r \mid \Theta)$

Sample $\mathcal{S}$ =

	A	B	C
d_1	1	0	1
d_2	0	1	0
d_3	0	0	1
d_4	1	0	1

- $P(d_1) = P_\Theta(+a, -b, +c)$
  $= P_\Theta(+a)\ P_\Theta(-b)\ P_\Theta(+c \mid +a, -b)$
  $= \theta_{+a}\ \theta_{-b}\ \theta_{+c|+a,-b}$
- $P(d_2) = P_\Theta(-a, +b, -c)$
  $= P_\Theta(-a)\ P_\Theta(+b)\ P_\Theta(-c \mid -a, +b)$
  $= \theta_{-a}\ \theta_{+b}\ \theta_{-c|-a,+b}$
- $P(\mathcal{S} \mid \Theta) = \Theta_{+a}^{2}\ \Theta_{-a}^{2}\ \Theta_{+b}^{1}\ \Theta_{-b}^{3}\ \Theta_{+c|+a,+b}^{0}\ \Theta_{+c|+a,-b}^{2} \cdots$
  $= \theta_{+a}^{N_{+a}}\ \theta_{-a}^{N_{-a}}\ \theta_{+b}^{N_{+b}}\ \theta_{-b}^{N_{-b}}\ \theta_{+c|+a,+b}^{N_{+c|+a,+b}}\ \theta_{+c|+a,-b}^{N_{+c|+a,-b}} \cdots$
  $= \prod_{ijk} \theta_{ijk}^{N_{ijk}}$

Example of Parameter θ_{ijk}



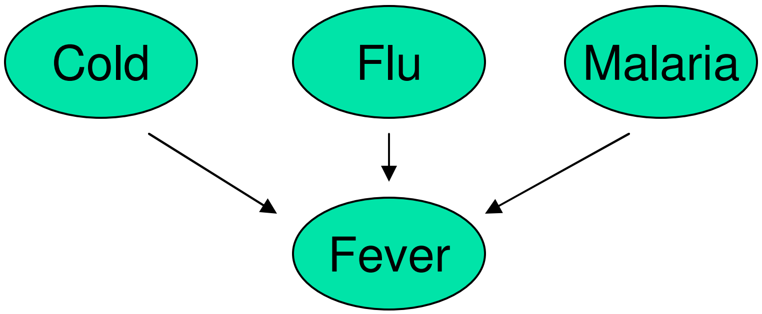
2^{nd} $\Downarrow$

4^{th} $\Rightarrow$

			$P(Fever = ? \mid$ Cold, Flu, Malaria$)$	
Cold	Flu	Malaria	True	False
F	F	F	θ_{111}	θ_{112}
F	F	T	θ_{121}	θ_{122}
F	T	F	θ_{131}	θ_{132}
F	T	T	θ_{141}	$\boxed{\theta_{142}}$
T	F	F	θ_{151}	θ_{152}
T	F	T	θ_{161}	θ_{162}
T	T	F	θ_{171}	θ_{172}
T	T	T	θ_{181}	θ_{182}

- $\theta_{ijk} = P(X_i = v_{ik} \mid Pa_i = pa_{ij})$
  - variable#1 -- here, "Fever"
  - 4th value of parents – [Cold=F, Flu=T, Malaria=T]

Example of Parameter N_{ijk}



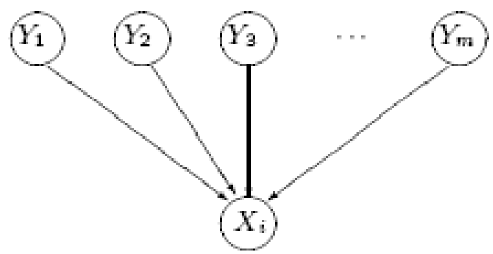
2nd ⇓

4th ⇒

			$P(Fever = ? \mid$ Cold, Flu, Malaria$)$	
Cold	Flu	Malaria	True	False
F	F	F	N_{111}	N_{112}
F	F	T	N_{121}	N_{122}
F	T	F	N_{131}	N_{132}
F	T	T	N_{141}	$\boxed{N_{142}}$
T	F	F	N_{151}	N_{152}
T	F	T	N_{161}	N_{162}
T	T	F	N_{171}	N_{172}
T	T	T	N_{181}	N_{182}

- N_{ijk} refers to …
  - variable#1 -- here, "Fever"
  - 4th value of parents – [Cold=F, Flu=T, Malaria=T]
  - 2nd value of Fever-node -- here, "Fever = FALSE"
- N_{ijk} is number of data-tuples where variable#i = its k^{th} value & parents(variable#i) = j^{th} value

Example of N_{ijk}, Θ_{ijk}



					$P(X_i = ? \mid Y_1, \ldots, Y_m)$				
	Y_1	Y_2	$\cdots$	Y_m	v_{i1}	$\cdots$	v_{ik}	$\cdots$	v_{ir_i}
	u_{11}	u_{21}	$\cdots$	u_{m1}	θ_{111}		θ_{11k}		θ_{11r_i}
	u_{11}	u_{21}	$\cdots$	u_{m2}	θ_{121}		θ_{12k}		θ_{12r_i}
	$\vdots$	$\vdots$	$\cdots$	$\vdots$					
$j^{th} \longrightarrow$	$u_{1\ell}$	$u_{2\ell'}$	$\cdots$	$u_{m\ell''}$			θ_{ijk}		
	$\vdots$	$\vdots$	$\cdots$	$\vdots$					
	u_{1r_1}	u_{2r_2}	$\cdots$	u_{mr_m}	θ_{1q_i1}		θ_{1q_ik}		$\theta_{1q_ir_i}$

- CPtable: $\theta_{ijk} = \hat{P}(X_i = v_{ik} \mid Pa_i = pa_{ij})$
- ...based on "Buckets"

	Y_1	Y_2	$\cdots$	Y_m	v_{i1}	$\cdots$	v_{ik}	$\cdots$	v_{ir_i}
	u_{11}	u_{21}	$\cdots$	u_{m1}	N_{111}		N_{11k}		N_{11r_i}
	u_{11}	u_{21}	$\cdots$	u_{m2}	N_{121}		N_{12k}		N_{12r_i}
	$\vdots$	$\vdots$	$\cdots$	$\vdots$					
$j^{th} \longrightarrow$	$u_{1\ell}$	$u_{2\ell'}$	$\cdots$	$u_{m\ell''}$			N_{ijk}		
	$\vdots$	$\vdots$	$\cdots$	$\vdots$					
	u_{1r_1}	u_{2r_2}	$\cdots$	u_{mr_m}	N_{1q_i1}		N_{1q_ik}		$N_{1q_ir_i}$

- N_{ijk} is number of data-tuples
  where variable#i = its k^{th} value
  and parents(variable#i) = j^{th} value

Task#1: Fixed Structure, Complete Tuples

- What are the ML values for Θ, given iid data $\mathcal{S} = \{ c_r \}$, ...

$$P(\mathcal{S} \mid \Theta) = \prod_{c \in \mathcal{S}} P(c \mid \Theta) = \prod_{c \in D} \prod_{[X_i = x_{ik}, Pa_i = pa_{ij}] \in c} \Theta_{ijk} =$$

$$\prod_{ijk} \Theta_{ijk}^{\ N_{ijk}} = \prod_{ij} \prod_{k} \Theta_{ijk}^{\ N_{ijk}}$$

- $\Theta^{(ML)} = \mathrm{argmax}_\Theta \{ P(\mathcal{S} \mid \Theta) \}$
$= \mathrm{argmax}_\Theta \{ \log P(\mathcal{S} \mid \Theta) \}$
$= \mathrm{argmax}_\Theta \{ \sum_{ij} \sum_k N_{ijk} \log \Theta_{ijk}) \}$

$\boxed{\forall ij \ \sum_k \Theta_{ijk} = 1}$

MLE Values

- $\Theta^{(ML)} = \text{argmax}_{\Theta} \{ \sum_{ij} \sum_k N_{ijk} \log \Theta_{ijk}) \}$

  $\boxed{\forall ij \ \sum_k \Theta_{ijk} = 1}$

- Notice $\theta_{ij.}$ is independent of $\theta_{rs.}$ when $i \neq r$ or $j \neq s$ …
  $\Rightarrow$ can solve each $\sum_k N_{ijk} \log \theta_{ijk}$ individually!

- For each $\sum_k N_{ijk} \log \theta_{ijk}$ … as $\sum_k \theta_{ijk} = 1$, optimum is

$$\boxed{\theta_{ijk} = \frac{N_{ijk}}{\sum_{k'} N_{ijk'}} = \frac{\#(X_i = v_{i,k} \ \& \ \mathbf{Pa}_i = \mathbf{pa}_{i,j})}{\#(\mathbf{Pa}_i = \mathbf{pa}_{i,j})}}$$

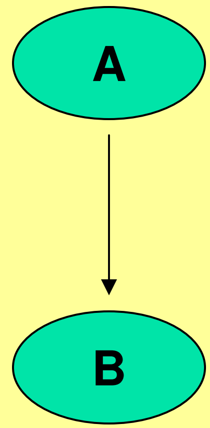
- Observed Frequency Estimates !
- Undefined if $\sum_k N_{ijk} = 0$ … $\#(\mathbf{Pa}_i = \mathbf{pa}_{i,j})=0$

Algorithm

ComputeMLE(graph $\mathcal{G}$, data $\mathcal{S}$):
return MLE parameters $[\theta_{ijk}]$

- Initialize $N_{ijk} \leftarrow 0$
- Walk thru data $\mathcal{S}$
  - Whenever see $[\ X_i = v_{ik},\ Pa_i = pa_{ij}]$,
    N_{ijk} += 1
- Return parameters: $\boxed{\theta_{ijk} = \frac{N_{ijk}}{\sum_r N_{ijr}}}$

Example



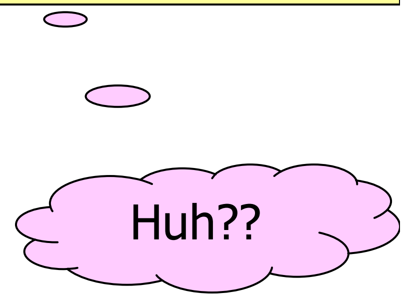
$$\Theta_A = \left[\frac{2}{2+0}, \frac{0}{2+0}\right]$$

$$\Theta_{B|+a} = \left[\frac{1}{1+1}, \frac{1}{1+1}\right]$$

$$\Theta_{B|-a} = \left[\frac{0}{0+0}, \frac{0}{0+0}\right]$$

- Buckets
  - $N_{+a} = \cancel{0}\ \cancel{1}\ 2$
  - $N_{-a} = 0$
  - $N_{+b|+a} = \cancel{0}\ 1$
  - $N_{-b|+a} = \cancel{0}\ 1$
  - $N_{+b|-a} = 0$
  - $N_{-b|-a} = 0$

A	B
+	+
+	−



Problems with MLE

- 0/0 issues
- Do you really believe 0% if 0 / 0+2 ?
- Which is better?
  - 3 heads, 2 tails $\theta =$ 3/(3+2) = 0.6
  - 30 heads, 20 tails $\theta =$ 30/(30+20) = 0.6
  - 3E23 heads, 2E23 tails $\theta =$ 3E23/(3E3+2E23) = 0.6
- What if you already know SOMETHING about the variable…

$\approx$ 50/50 …

Bayesian Learning

- Use Bayes rule:

$$\underbrace{P(\theta \mid \mathcal{D})}_{\text{posterior}} = \frac{\overbrace{P(\mathcal{D} \mid \theta)}^{\text{likelihood}}\overbrace{P(\theta)}^{\text{prior}}}{P(\mathcal{D})}$$

- Or equivalently:

$$P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta)$$

Bayesian Learning

$$P(\theta \mid \mathcal{D}) \;\; \propto \;\; P(\mathcal{D} \mid \theta) P(\theta)$$

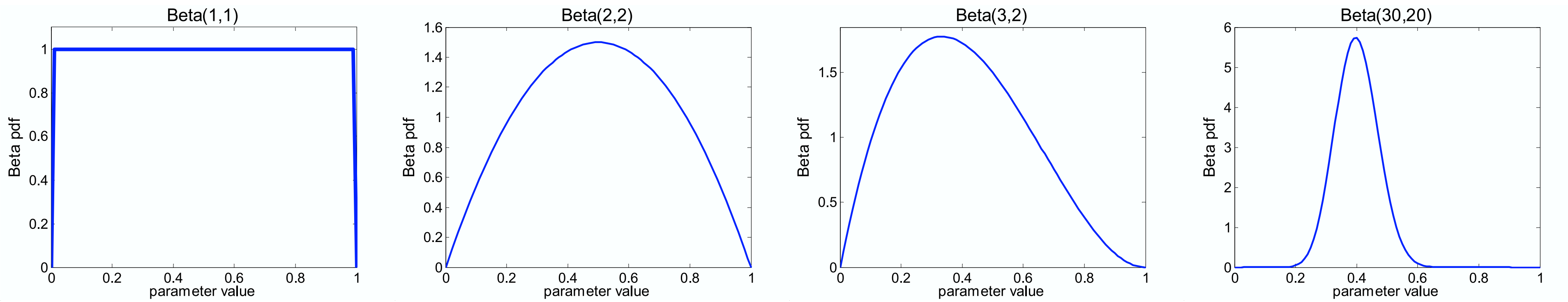
posterior *likelihood* *prior*

- Likelihood function is simply Binomial:

$$P(\mathcal{D} \mid \theta) = \theta^{m_H}(1-\theta)^{m_T}$$

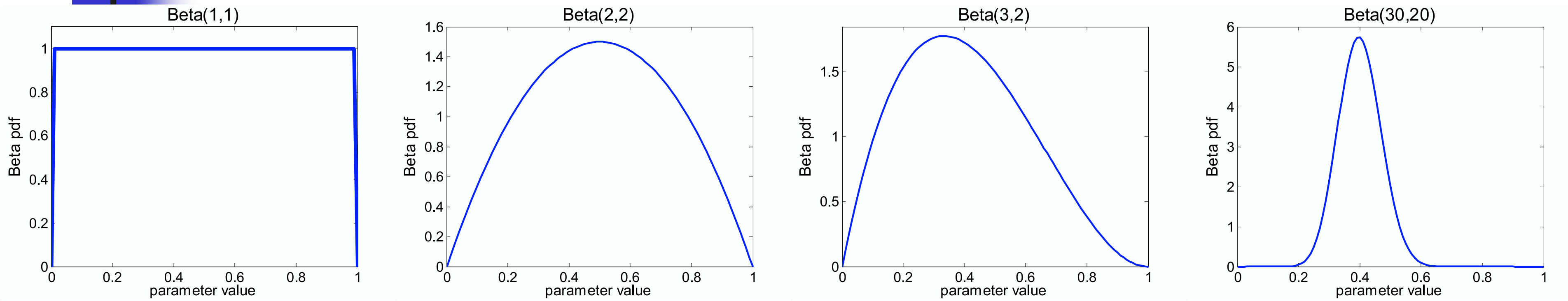
- What about prior?
  - Represent expert knowledge
  - Simple posterior form
- Conjugate priors:
  - Closed-form representation of posterior (more details soon)
  - **For Binomial, conjugate prior is Beta distribution**

Beta Prior Distribution – P(θ)



- Prior: $$P(\theta) = \frac{\theta^{\alpha_H-1}(1-\theta)^{\alpha_T-1}}{B(\alpha_H,\alpha_T)} \sim Beta(\alpha_H,\alpha_T)$$
- Likelihood function: $P(\mathcal{D} \mid \theta) = \theta^{m_H}(1-\theta)^{m_T}$
- Given X ~ Beta(a, b) :
  - Mean: a/(a + b)
  - Unimodal if a,b>1… here mode: (a-1) / (a+b-2)
  - Variance: $a \times b / [(a+b)^2 (a+b-1)]$

Posterior distribution… from Beta



Prior P(θ)

Likelihood P(D|θ)

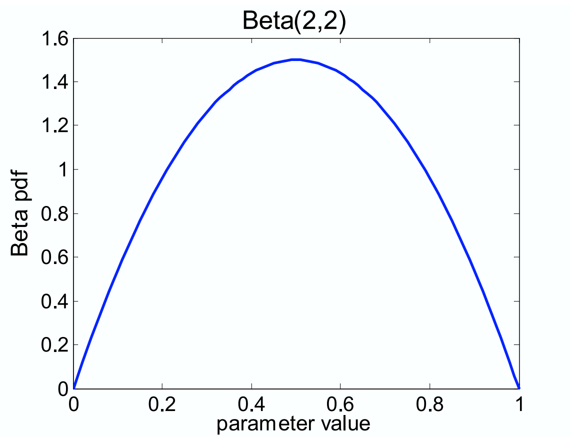
$$
\begin{aligned}
P(\theta \mid \mathcal{D}) &\propto P(\theta)\ P(\mathcal{D} \mid \theta) \\
&= \Theta^{\alpha_H - 1}(1-\Theta)^{\alpha_T - 1} \times \Theta^{m_H}(1-\Theta)^{m_T} \\
&= \Theta^{\alpha_H + m_H - 1}(1-\Theta)^{\alpha_T + m_T - 1} \\
&\sim \text{Beta}(\alpha_H + m_H,\ \alpha_T + m_T)
\end{aligned}
$$

Same form! Conjugate!

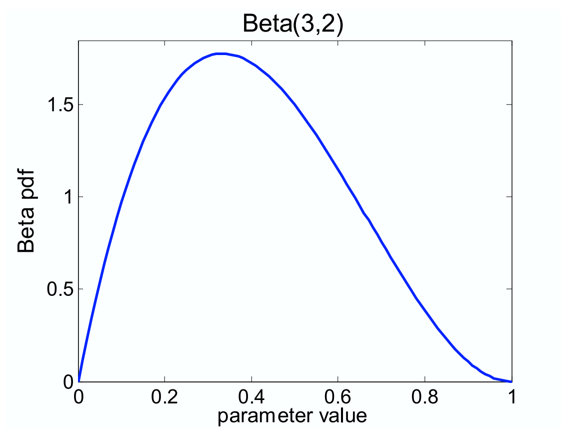
Posterior Distribution

- Prior: $\theta \sim \mathrm{Beta}(\alpha_H, \alpha_T)$
- Data $\mathcal{S}$: m_H heads, m_T tails
- Posterior distribution:

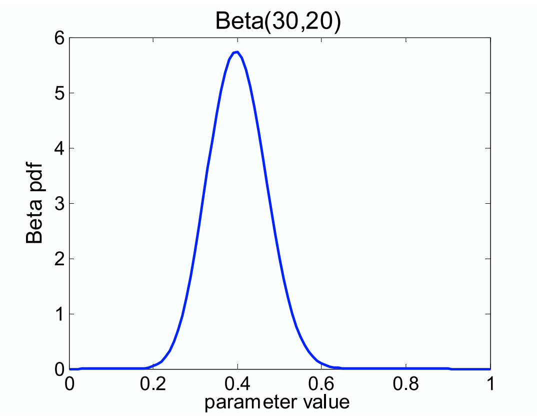
  $\theta \mid \mathcal{S} \sim \mathrm{Beta}(m_H + \alpha_H,\ m_T + \alpha_T)$



Prior

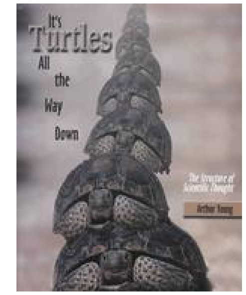


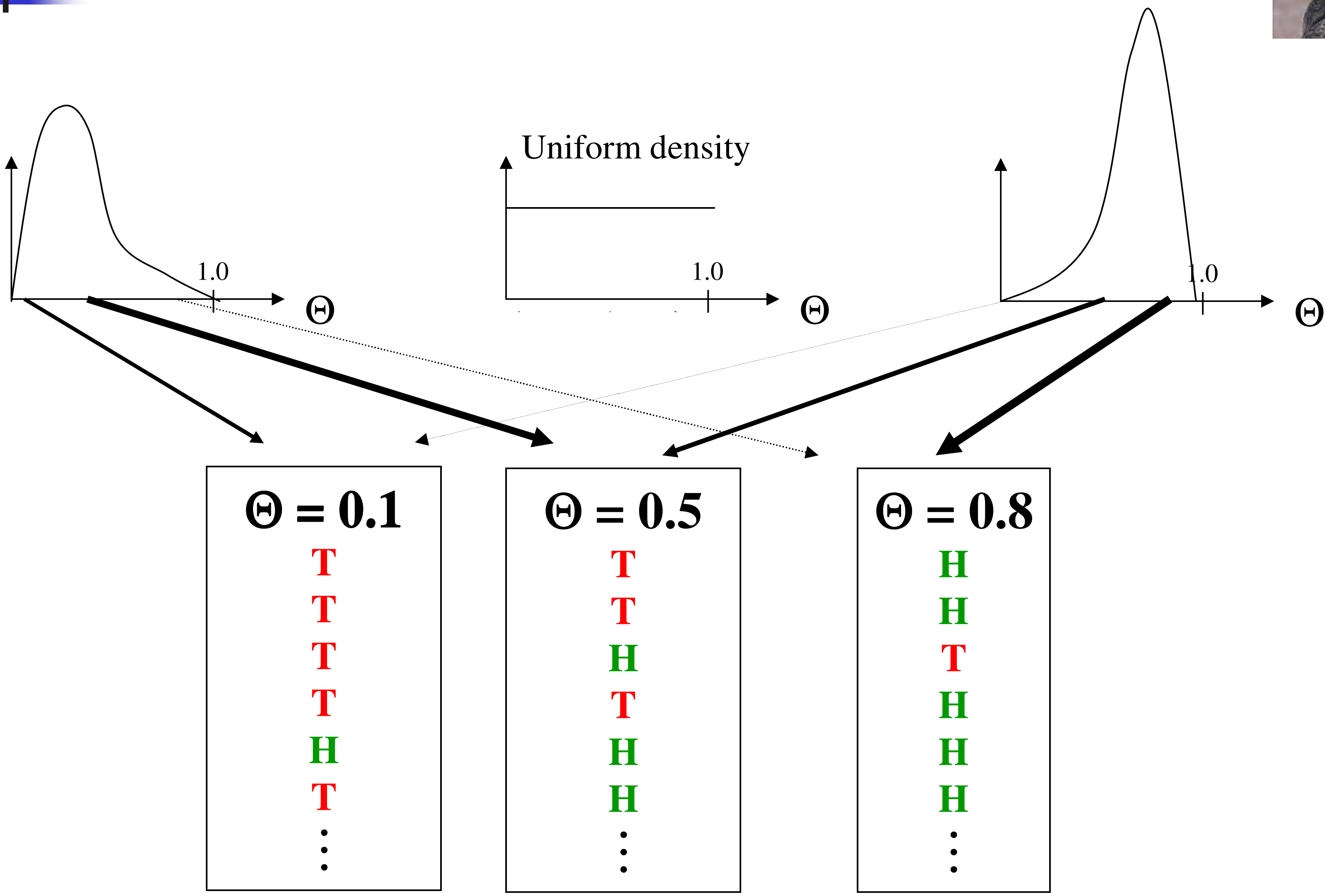
+ observe 1 head



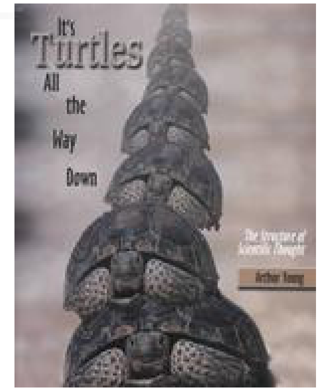
+ observe 27 more heads; 18 tails

Two (related) Distributions: Parameter, Instances





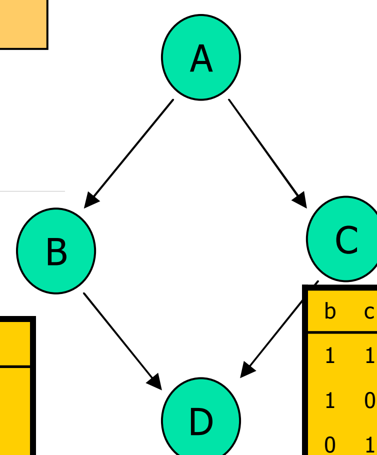
Distribution over Parameter



- What is "real" value of $\theta_{A=1}$?
- If …
  - uncertainty in expert opinion
  - limited training data

only a distribution!

$$\theta_{A=1} \sim \boxed{\text{Beta}(\ 4,\ 6\ )}$$

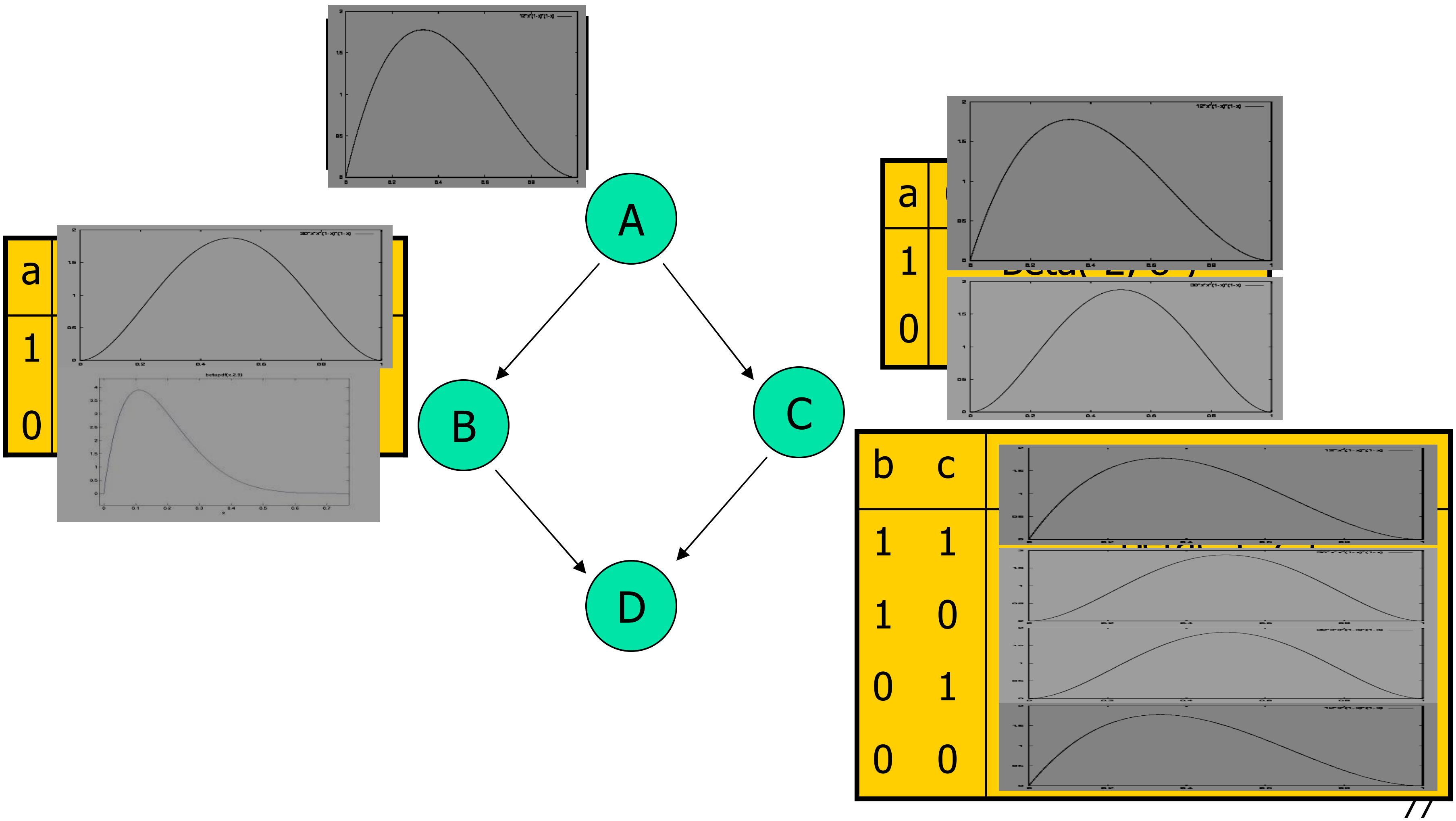


a	$\theta_{C=1\|A=a}$	$\theta_{C=0\|A=a}$
1	0.200	0.800
0	0.367	0.633

a	$\theta_{B=1\|A=a}$	$\theta_{B=0\|A=a}$
1	0.325	0.675
0	0.440	0.550

b	c	$\theta_{D=1\|B=b,C=c}$	$\theta_{D=0\|B=b,C=c}$
1	1	0.300	0.700
1	0	0.333	0.667
0	1	0.250	0.750
0	0	0.450	0.550

Distribution over Parameters



Beta Distribution

- Model row-parameter

  $\theta_{B|a=1} = \langle\, \theta_{b=0|a=1},\ \ \theta_{b=1|a=1} \,\rangle$

  as *Beta distribution*

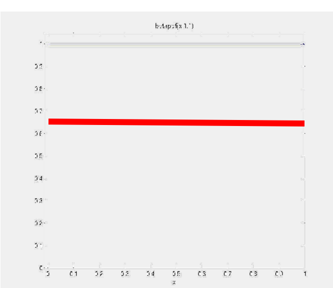
- $\theta_{B|A=1} = \langle \theta_{B=0|A=1},\ \ \theta_{B=1|A=1} \rangle \sim$ Beta(1, 1)

  kinda like seeing 2 instances with $\langle A=1 \rangle$:

1 with $\langle A=1, B=0 \rangle$

1 with $\langle A=1, B=1 \rangle$

A	B	C	D
1	0	0	1
1	1	1	1
0	0	1	1
⋮	⋮	⋮	⋮



Beta Distribution, II

- $\theta_{B|A=1} = \langle \theta_{B=0|A=1},\ \theta_{B=1|A=1} \rangle \sim$ Beta(1, 1)

$$\Rightarrow \quad \boxed{E[\theta_{B=0|A=1}] = \widehat{\theta}_{-b|+a} = \frac{1}{1+1} = 0.5}$$

- Now… observe data S :

A	B	C	E
1	1	0	1
1	1	1	1
1	**0**	1	0
1	**0**	1	0
1	**0**	0	0
1	**0**	0	1
0	0	0	1
⋮	⋮	⋮	⋮

6 "⟨*A=1*⟩"

2 "⟨*A=1, B=1*⟩"s

4 "⟨*A=1,B=0*⟩"s

Beta Distribution, III

- $\theta_{B|A=1} = \langle \theta_{B=0|A=1},\ \theta_{B=1|A=1} \rangle \sim$ Beta(1, 1)

$\Rightarrow$ $$E[\theta_{B=1|A=1}] = \widehat{\theta}_{+b|+a} = \frac{1}{1+1} = 0.5$$

- Then observe data $\mathcal{D}$
  - *2* $\langle A=1, B=1 \rangle$
  - *4* $\langle A=1, B=0 \rangle$

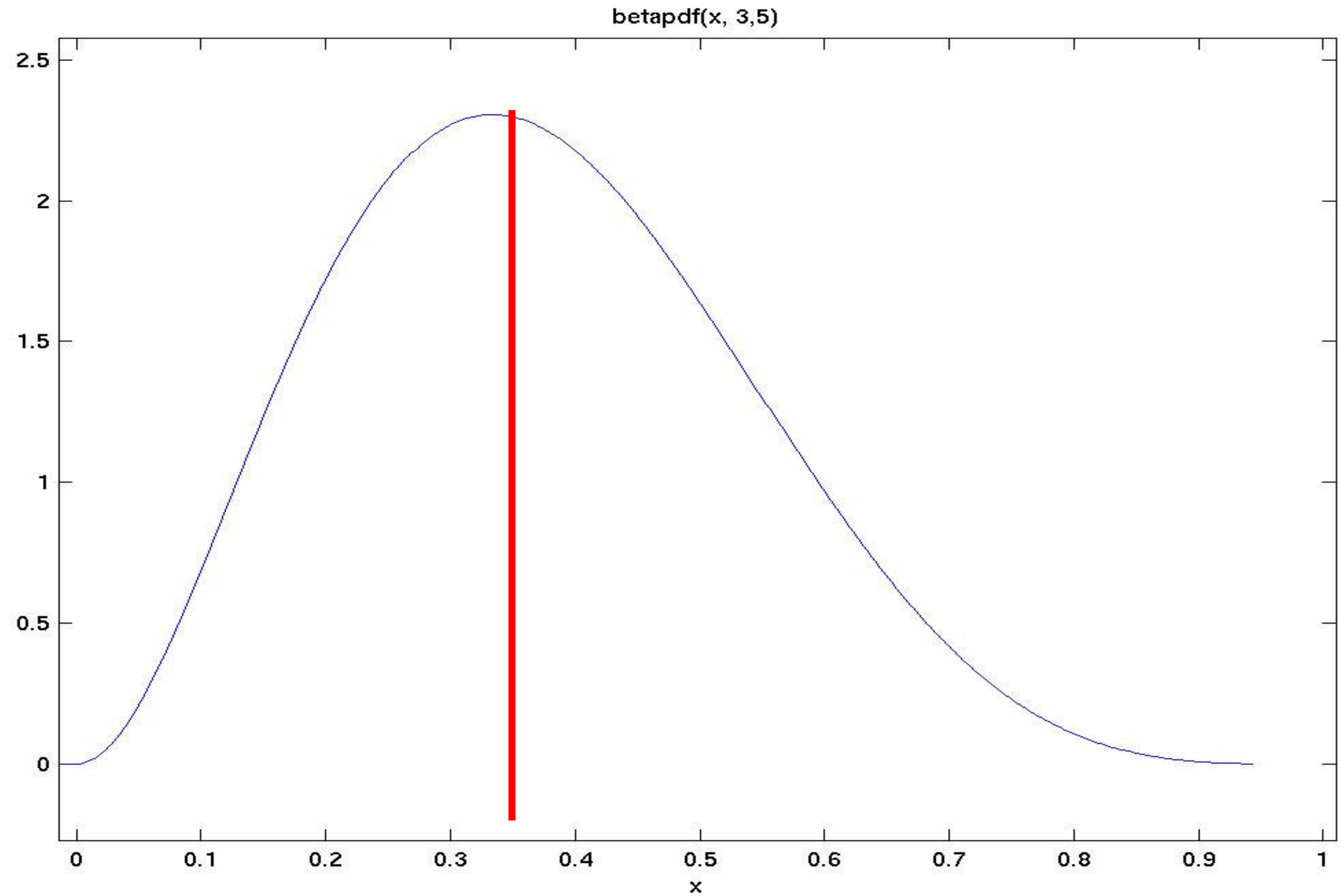
A	B	C	E
1	1	0	1
1	1	1	1
1	0	1	0
1	0	1	0
1	0	0	0
1	0	0	1
0	0	0	1
⋮	⋮	⋮	⋮

- *New distribution is*

  $\theta'_{B|A=1} \sim$ Beta(1+2, 1+4) = Beta(3, 5)

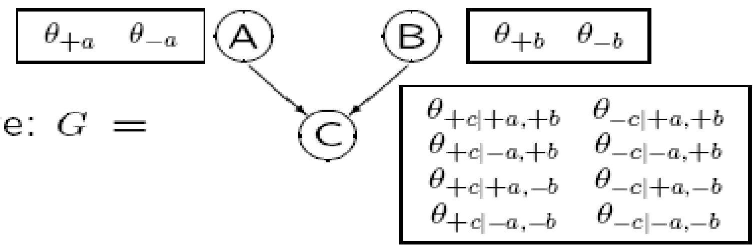
$\Rightarrow$ $$E[\theta_{B=1|A=1} \mid S] = \widehat{\theta}_{+b|+a} \mid S = \frac{3}{3+5} = 0.375$$

$\theta_{B|+a}$ ~ Beta(3,5) Distribution

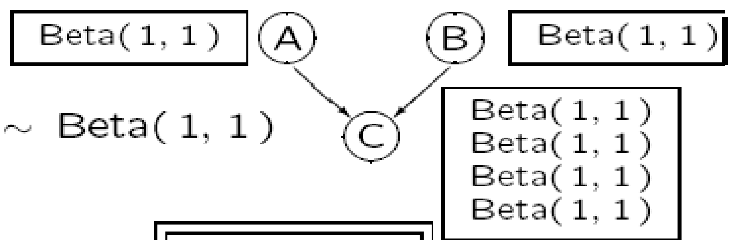


Posterior Distribution of Θ

Given: Structure: G =



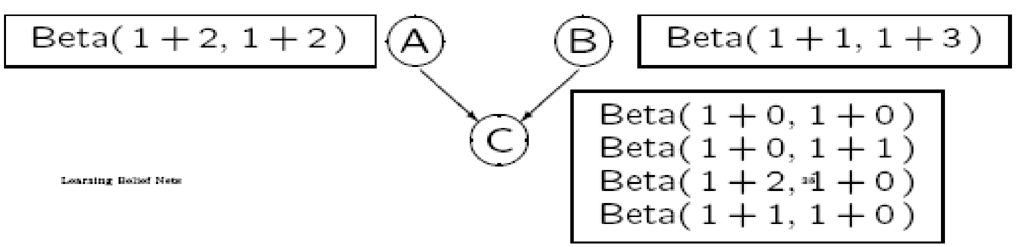
where $\Theta_{X|y} \sim$ Beta(1, 1)



- Given sample S =

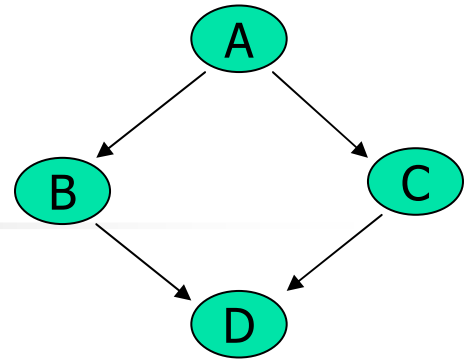
	A	B	C
d_1	1	0	1
d_2	0	1	0
d_3	0	0	1
d_4	1	0	1

Posterior distribution is. . .



Learning Belief Nets

Posterior Distribution



- Initially: $P(X_i \mid pa_{ij})$ …
  $\theta_{ij} \sim Dir(\alpha_{ij1}, \ldots, \alpha_{ijr})$
- Data $\mathcal{S}$ includes
  N_{ijk} examples including $[X_i = v_{ik}, \mathbf{Pa_i} = pa_{ij}]$
- Posterior
  $\theta_{ij} \mid \mathcal{S} \sim Dir(\alpha_{ij1} + N_{ij1}, \ldots, \alpha_{ijr} + N_{ijr})$
- Expected value

$$E[\theta_{ijk}] = \frac{\alpha_{ijk} + N_{ijk}}{\sum_r \alpha_{ijr} + N_{ijr}}$$

- Compare to Frequentist:

$$\hat{\theta}_{ijk} = \frac{N_{ijk}}{\sum_r N_{ijr}}$$

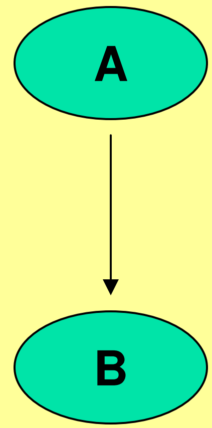
Algorithm

ComputePosterior(graph $\mathcal{G}$, data $\mathcal{S}$, priors $[\alpha_{ijk}]$):
return posterior parameters $[N_{ijk}]$

- Initialize $N_{ijk} \leftarrow \alpha_{ijk}$
- Walk thru data $\mathcal{S}$
  - Whenever see $[\ X_i = v_{ik},\ Pa_i = pa_{ij}]$,
    N_{ijk} += 1
- Set parameters:
  $\theta_{ij} \mid \mathcal{S} \sim Dir(\ N_{ij1}, \ldots, N_{ijr})$
- If want expected value:

$$E[\theta_{ijk}] = \frac{N_{ijk}}{\sum_r N_{ijr}}$$

Example



$\theta_A \sim \mathrm{Beta}(\alpha_{+a}, \alpha_{-a})$

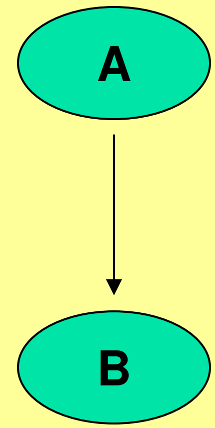
$\theta_{B|+a} \sim \mathrm{Beta}(\alpha_{+b|+a}, \alpha_{-b|+a})$

$\theta_{B|-a} \sim \mathrm{Beta}(\alpha_{+b|-a}, \alpha_{-b|-a})$

- Buckets
  - $N_{+a} := \alpha_{+a}$
  - $N_{-a} := \alpha_{-a}$
  - $N_{+b|+a} := \alpha_{+b|+a}$
  - $N_{-b|+a} := \alpha_{-b|+a}$
  - $N_{+b|-a} := \alpha_{+b|-a}$
  - $N_{-b|-a} := \alpha_{-b|-a}$

A	B
+	+
+	−

Example



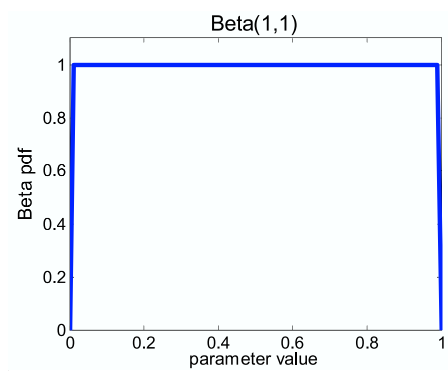
$\theta_A \sim \text{Beta}(3, 1)$

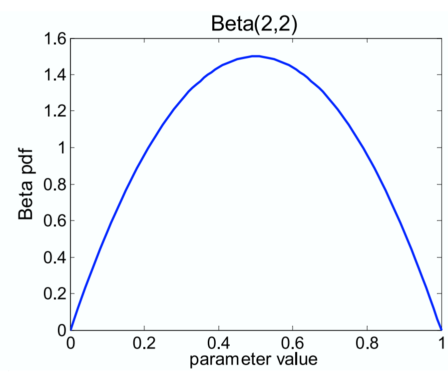
$\theta_{B|+a} \sim \text{Beta}(2, 2)$

$\theta_{B|-a} \sim \text{Beta}(2, 7)$

- Buckets
  - $N_{+a} := \cancel{1}\ \cancel{2}\ 3$
  - $N_{-a} := 1$
  - $N_{+b|+a} := \cancel{1}\ 2$
  - $N_{-b|+a} := \cancel{1}\ 2$
  - $N_{+b|-a} := 2$
  - $N_{-b|-a} := 7$

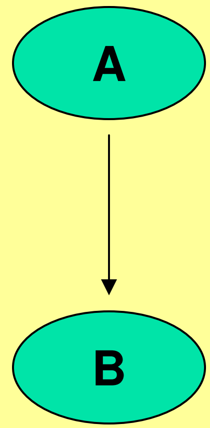
A	B
+	+
+	−





Example

If you want POINT estimates…



$$\hat{\Theta}_A = \left[\frac{3}{3+1}, \frac{1}{3+1}\right]$$

$$\hat{\Theta}_{B|+a} = \left[\frac{2}{2+2}, \frac{2}{2+2}\right]$$

$$\hat{\Theta}_{B|-a} = \left[\frac{2}{2+7}, \frac{7}{2+7}\right]$$

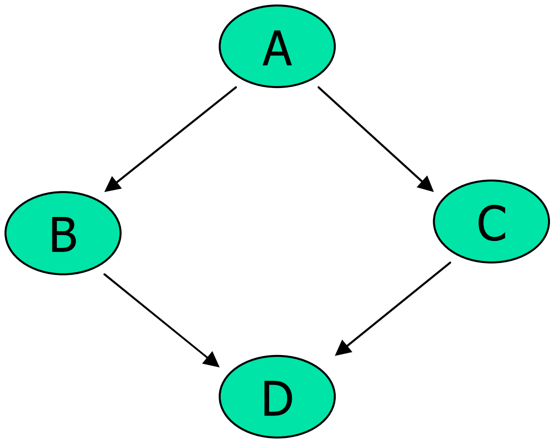
- Buckets
  - $N_{+a} := \cancel{1}\ \cancel{2}\ 3$
  - $N_{-a} := 1$
  - $N_{+b|+a} := \cancel{1}\ 2$
  - $N_{-b|+a} := \cancel{1}\ 2$
  - $N_{+b|-a} := 2$
  - $N_{-b|-a} := 7$

A	B
+	+
+	−

Note: no 0/0 issues!

In general, should initialize N_{ijk} to α_{ijk} … called "pseudo-counts"

Answer to a Query…



- Response to query

  $$P_\Theta(\ C{=}c \mid \mathbf{E}{=}\mathbf{e}\ )$$

  is function of parameters Θ
- Eg…

$$P_\Theta(A=1 \mid B=1, C=1) \;=\; \frac{\theta_{A=1}\,\theta_{B=1|A=1}\,\theta_{C=1|A=1}}{\sum_a \theta_{A=a}\,\theta_{B=1|A=a}\,\theta_{C=1|A=a}}$$

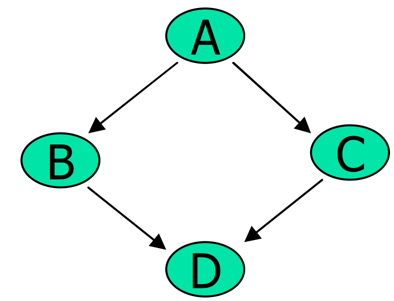
What is $P_\Theta(C=c \mid \mathbf{E}=\mathbf{e})$?

- $P_\Theta(C=c \mid \mathbf{E}=\mathbf{e})$ depends on Θ
- As Θ is r.v., so is response

  $q(\Theta) = P_\Theta(C=c \mid \mathbf{E}=\mathbf{e})$
- Properties of $q(\Theta)$
  - within [0,1]
  - Mean

$$E[\, q(\Theta) \,] = \int_\Theta q(\Theta)\, P(\Theta)\, d\Theta$$

How to compute $E[\, P_\Theta(\, C{=}c \mid \mathbf{E{=}e} \,)\,]$?

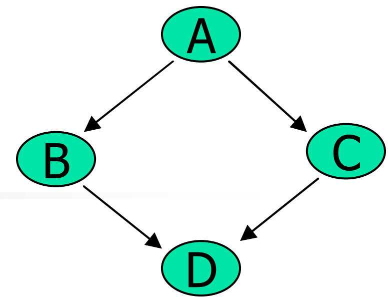


$$q(\Theta) = P_\Theta(A=1 \mid B=1, C=1) = \frac{\theta_{A=1}\,\theta_{B=1|A=1}\,\theta_{C=1|A=1}}{\sum_a \theta_{A=a}\,\theta_{B=1|A=a}\,\theta_{C=1|A=a}}$$

- Draw R samples $\Theta^{(i)}$ from $P(\Theta)$
  - $\Theta_A \sim Be(3,7)$, $\Theta_{B|+a} \sim Be(1,4)$, ...
  - $\Theta_A^{(1)} = [0.29, 0.71]$; $\Theta_{B|+a}^{(1)} = [0.18, 0.82]$; ...
    $q(\Theta^{(1)}) = 0.57$
  - $\Theta_A^{(2)} = [0.32, 0.68]$; $\Theta_{B|+a}^{(2)} = [0.23, 0.77]$; ...
    $q(\Theta^{(2)}) = 0.61$
  - ...
- Let $q^{(R)} = 1/R \sum_i q(\, \Theta^{(i)}\,)$
- As $R \to \infty$, $q^{(R)} \to E[q]$

But … easier approach:

Predictive Distribution



- If $q(\theta)$ is UNCONDITIONAL query,

  $q(\Theta) = P_\Theta(+a, +b, -c) = \Theta_{+a}\, \Theta_{+b|+a}\, \Theta_{-c|+a}$

$$\boxed{\hat{q} = E[\ q(\Theta)\ ] \ = \ q(E_\Theta[\Theta]) \ = \ q(\hat{\Theta}) \ !}$$

- $BN^{\mathcal{D}} = [\mathcal{G}, \Theta^{\mathcal{D}}]$ with $\theta^{\mathcal{D}} = \left\{ \frac{N_{ijk}+1}{\sum_k (N_{ijk}+1)} \right\}$

  Compute $E[\ q(\theta)\ ]$ by using just $BN^{\mathcal{D}}$!

  $\Rightarrow$ get Model-Averaging for free!
- More complicated for Conditional Queries!

Summary: Parameter Learning

- MLE:
  - score decomposes according to CPTs
  - optimize each CPT separately
- Bayesian parameter learning:
  - motivation for Bayesian approach
  - Bayesian prediction
  - conjugate priors, equivalent sample size
  - Bayesian learning ⇒ smoothing
- Bayesian learning for BN parameters
  - Global parameter independence
  - Decomposition of prediction according to CPTs
  - Decomposition within a CPT
  - Predictive distribution – model averaging, for free!

Complete Data…

Outline

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

Skip

#2: Known structure, Missing data

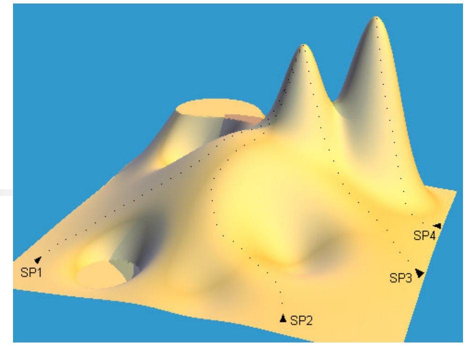
- To find good Θ, need to compute $P(\Theta, \mathcal{D} \mid \mathcal{G})$
- Easy if ..

$$S = \left\{ \begin{array}{lllll} c_1 : & \langle \boxed{\phantom{c_{11}}} & \cdots & c_{1N} & \rangle \\ c_2 : & \langle c_{21} & \cdots & \boxed{\phantom{c_{2N}}} & \rangle \\ \vdots & \langle \;\vdots & c_{ij} & \;\vdots & \rangle \\ c_m : & \langle c_{m1} & \cdots & c_{mN} & \rangle \end{array} \right\}$$

~~complete~~ incomplete

- What if S is incomplete
  - Some $c_{ij} = *$
  - "Hidden variables" (X_K never seen: $c_{iK} = * \;\forall$ i)
- Here:
  - Given fixed structure
  - Missing (Completely) At Random:
    Omission not correlated with value, etc.
- Approaches:
  - Gradient Ascent, EM, Gibbs sampling, ...

Gradient Ascent



- Want to maximize likelihood
  - $\theta^{(MLE)} = \text{argmax}_\theta\, L(\theta : S)$
- Unfortunately…
  - $L(\theta : S)$ is nasty, non-linear, multimodal fn
  - So…
- Gradient-Ascent
  - … 1st-order Taylor series

$$f_{obj}(\theta^*) \approx f_{obj}(\theta^0) + (\theta - \theta^0)^T \nabla f_{obj}($$

Need derivative!

Procedure Gradient-Ascent (
  θ^1, // Initial starting point
  f_{obj}, // Function to be optimized
  δ // Convergence threshold
)

```
t ← 1
do
    θ^{t+1} ← θ^t + η∇f_obj(θ^t)
    t ← t + 1
while ||θ^t − θ^{t−1}|| > δ
return (θ^t)
```

Gradient Ascent [APN]

View: $P_\Theta(S) = P(S \mid \Theta, \text{G})$ as fn of Θ

$$\frac{\partial \ln P_\Theta(S)}{\partial \theta_{ijk}} = \sum_{\ell=1}^{m} \frac{\partial \ln P_\Theta(c_\ell)}{\partial \theta_{ijk}} = \sum_{\ell=1}^{m} \frac{\partial P_\Theta(c_\ell)/\partial \theta_{ijk}}{P_\Theta(c_\ell)}$$

$$\frac{\partial P_\Theta(c_\ell)/\partial \theta_{ijk}}{P_\Theta(c_\ell)} = \frac{P_\Theta(c_\ell \mid v_{ik}, \mathbf{pa}_{ij})\, P_\Theta(\mathbf{pa}_{ij})}{P_\Theta(c_\ell)} = \frac{P_\Theta(v_{ik}, \mathbf{pa}_{ij} \mid c_\ell)}{\theta_{ijk}}$$

Alg: fn Basic-APN(BN = $\langle$ G, Θ $\rangle$, $\mathcal{D}$): (modified) CPtables
inputs: BN, a Belief net with CPT entries
$\mathcal{D}$, a set of data cases
repeat until $\Delta\Theta \approx 0$
$\Delta\Theta \leftarrow 0$
for each $c_r \in \mathcal{D}$
Set evidence in BN to c_r
For each X_i w/ value v_{ik}, parents w/ j^{th} value pa_{ij}
$\Delta\Theta_{ijk}$ += $P(v_{ik}, pa_{ij} \mid c_r) / \theta_{ijk}$
Θ += $\alpha\, \Delta\Theta$
$\Theta \leftarrow$ project Θ onto constraint region
return(Θ)

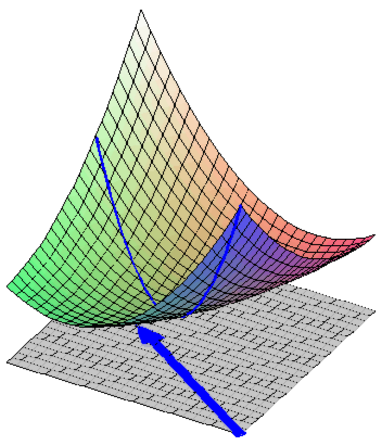
Note: Computed $P(v_{ik}, pa_{ij} \mid c_r)$ to deal with c_r
$\Rightarrow$ can "piggyback" computation

Issues with Gradient Ascent

- Constraints
  - $\Theta_{ijk} \in [0,1]$
  - $\sum_r \Theta_{ijr} = 1$
  - But … Θ_{ijk} += $\alpha\, \Delta\Theta_{ijk}$ could violate
  - Use $\Theta_{ijk} = \exp(\lambda_{ijk}) / \sum_r \exp(\lambda_{ijr})$
  - Find best λ_{ijk} … unconstrained …

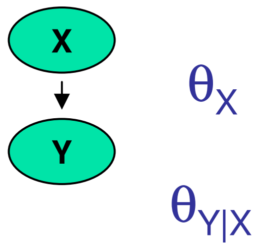
- Lots of Tricks for efficient ascent
  - Line Search
  - Conjugate Gradient
  - …

  Take Cmput551, or optimization



Expectation Maximization (EM)

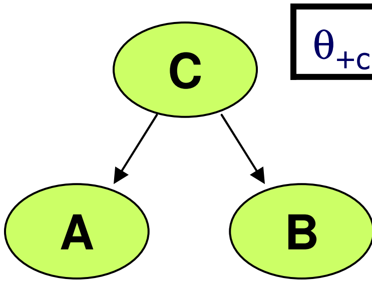
- EM is designed to find most likely θ, given incomplete data !
- Recall simple Maximization needs counts: #(+x, +y), …
- But is instance [?, +y] in
  … #(+x, +y)? … #(-x, +y)?



- Why not put it in BOTH… fractionally ?
  - What is weight of #(+x, +y)?
  - $P_\theta(+x \mid +y)$, based on current value of θ

EM Approach – E Step

θ_{+c}	θ_{-c}



$\theta_{+a\|+c}$	$\theta_{-a\|+c}$
$\theta_{+a\|-c}$	$\theta_{-a\|-c}$

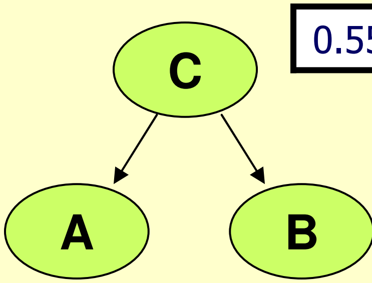
$\theta_{+b\|+c}$	$\theta_{-b\|+c}$
$\theta_{+b\|-c}$	$\theta_{-b\|-c}$

Sample S =

A	B	C
0	0	1
*	1	0
0	*	1
*	*	1

Guess initial values θ^0

0.55	0.45



0.8	0.2
0.3	0.7

0.9	0.1
0.4	0.6

Set $S^{(0)}$ =

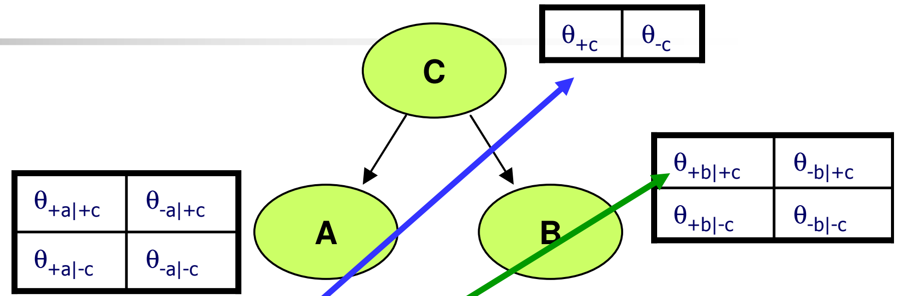
A	B	C	
0	0	1	1.0
0	1	0	0.7
1	1	0	0.3
0	0	1	0.1
0	1	1	0.9
0	0	1	0.7×0.1
0	1	1	0.7×0.9
1	0	1	0.3×0.1
1	1	1	0.3×0.9

EM Approach – M Step

•Use fractional data:

$S^{(0)} =$

A	B	C	
0	0	1	1.0
0	1	0	0.7
1	1	0	0.3
0	0	1	0.1
0	1	1	0.9
0	0	1	0.7×0.1
0	1	1	0.7×0.9
1	0	1	0.3×0.1
1	1	1	0.3×0.9



•New estimates:

$$\hat{\theta}^{(1)}_{+a|+c} = \frac{\#(+a,+c)}{\#(+c)} = \frac{(0.3\times0.1)+(0.3\times0.9)}{1+0.1+0.9+(0.7\times0.1)+(0.7\times0.9)+(0.3\times0.1)+(0.3\times0.9)} = 0.1$$

$$\hat{\theta}^{(1)}_{+b|+c} = \frac{\#(+b,+c)}{\#(+c)} = \frac{0.1+(0.7\times0.9)+(0.3\times0.9)}{3} = 0.33$$

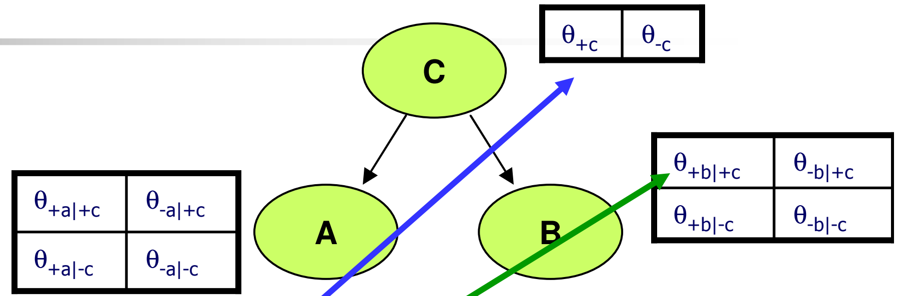
$$\hat{\theta}^{(1)}_{+c} = \frac{\#(+c)}{\#(\{\})} = \frac{1.0+(1.0)+(1.0)}{4} = 0.75$$

EM Approach – M Step

•Use fractional data:

$S^{(0)}$ =

A	B	C	
0	0	1	
0	1	0	
1	1	0	
0	0	1	
0	1	1	
0	0	1	
0	1	1	
1	0	1	
1	1	1	



•New estimates:

$$\hat{\theta}^{(1)}_{+a|+c} = \frac{\#(+a,+c)}{\#(+c)} = \frac{(0.3\times0.1)+(0.3\times0.9)}{1+0.1+0.9+(0.7\times0.1)+(0.7\times0.9)+(0.3\times0.1)+(0.3\times0.9)} = 0.1$$

$$\hat{\theta}^{(1)}_{+b|+c} = \frac{\#(+b,\ldots}{\#(+c\ldots}$$

$$\hat{\theta}^{(1)}_{+c} = \frac{\#(+c)}{\#(\{\})} = \ldots$$

Then

- **E-step**: re-estimate distributions over the missing values based on these new $\theta^{(1)}$ values
- **M-step**: compute new $\theta^{(2)}$ values, using statistics based on these new distribution

EM Steps

- **E step**:
  - Given parameters $\theta^{(t)}$
  - find probability of each missing value
    - … so get $E_{\theta(t)}[\ N_{ijk}\ ]$
- **M step**:
  - Given completed (fractional) data
    - based on $E_{\theta(t)}[\ N_{ijk}\ ]$
  - find max-likely parameters $\theta^{(t+1)}$

EM Approach

- Assign $\Theta^{(0)} = \{\theta_{ijk}^{(0)}\}$ randomly.

- Iteratively, $k = 0, \ldots$

  E step: Compute EXPECTED value of N_{ijk}, given $\langle \mathsf{G}, \Theta^k \rangle$

$$\hat{N}_{ijk} = E_{P(\,x \,|\, S, \Theta^k, \mathsf{G}\,)}(N_{ijk}) = \sum_{c_\ell \in S} P(\, x_i^k, \mathbf{pa}_i^j \,|\, c_\ell, \Theta^k, S\,)$$

  M step: Update values of Θ^{k+1}, based on $\hat{N}_{ijk}$

$$\theta_{ijk}^{k+1} = \frac{\hat{N}_{ijk} + 0}{\sum_{k=1}^{r_i} (\hat{N}_{ijk} + 0)}$$

  . . . until $\| \Theta^{k+1} - \Theta^k \| \approx 0.$

- Return Θ^k

1. This is ML computation; MAP is similar
   "0" $\to \alpha_{ijk}$
2. Finds local optimum
3. Used for HMM
4. Views each tuple with k "$*$"s as $O(2^k)$ partial-tuples

Facts about EM …

- Always converges
- Always improve likelihood
  - $L(\theta^{(t+1)} : \mathcal{S}) > L(\theta^{(t)} : \mathcal{S})$
  - … except at stationary points…

- For CPtable for Belief net:
  - Need to perform general BN inference
  - Use Click-tree or ClusterGraph
    … just needs one pass
    (as N_{ijk} depends on node+parents)

Gibbs Sampling

- Let $S^{(0)}$ be COMPLETED version of S,
  randomly filling-in each missing c_{ij}

  Let $d_{ij}^{(0)} = c_{ij}$
  If $c_{ij} = *$, then $d_{ij}^{(0)} = \text{Random[Domain}(X_i)\text{]}$

- For $k = 0..$
  - Compute $\Theta^{(k)}$ from $S^{(k)}$ [frequencies]
  - Form $S^{(k+1)}$ by...
    - $d_{ij}^{k+1} = c_{ij}$
    - If $c_{ij} = *$ then
      Let d_{ij}^{k+1} be random value for X_i,
      based on current distr Θ^k over $Z - X_i$

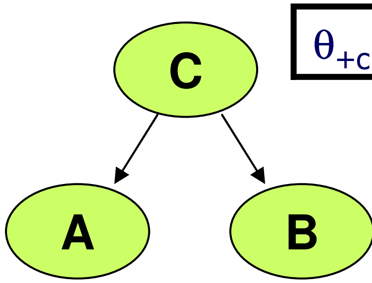
- Return average of these $\Theta^{(k)}$'s

Note: As $\Theta^{(k)}$ based on COMPLETE DATA $S^{(k)}$
$\Rightarrow \Theta^{(k)}$ can be computed efficiently!

"Multiple Imputation"

Gibbs Sampling – Example

θ_{+c}	θ_{-c}



$\theta_{+a\|+c}$	$\theta_{-a\|+c}$
$\theta_{+a\|-c}$	$\theta_{-a\|-c}$

$\theta_{+b\|+c}$	$\theta_{-b\|+c}$
$\theta_{+b\|-c}$	$\theta_{-b\|-c}$

New $S^{(1)}$ =

A	B	C
0	0	1
0	1	0
0	1	1
1	1	1

Flip 0.3-coin:

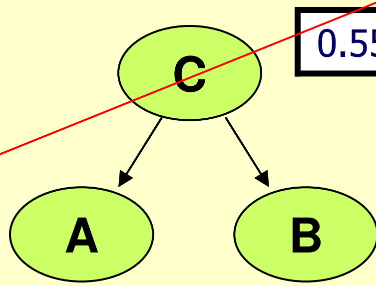
Flip 0.9-coin:

Flip 0.8-coin:

Flip 0.9-coin:

Guess initial values θ^0

0.55	0.45



0.8	0.2
0.3	0.7

0.9	0.1
0.4	0.6

Then

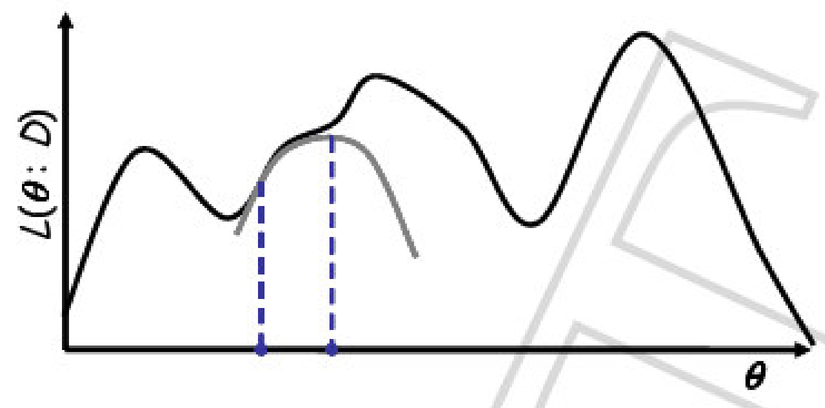
- Use $S^{(1)}$ to get new $\theta^{(2)}$ parameters
- Form new $S^{(2)}$ by drawing new values from $\theta^{(2)}$

Gibbs Sampling (con't)

- Algorithm: Repeat
  - Given COMPLETE data $S^{(i)}$, compute new ML values for $\{\theta_{ijk}^{(i+1)}\}$
  - Using NEW parameters, impute (new) missing values $S^{(i+1)}$
- Q: What to return?
  AVERAGE over separated $\Theta^{(i)}$'s
  - eg, $\Theta^{(500)}$, $\Theta^{(600)}$, $\Theta^{(700)}$, …
- Q: When to stop?
  When distribution over $\Theta^{(i)}$s have converged
- Comparison: Gibbs vs EM
  - + EM "splits" each instance
    …into 2^k parts if k *'s
  - – EM knows when it is done, and what to return

General Issues

- All alg's are heuristic…
  - Starting values θ
  - Stopping criteria
  - Escaping local maxima



- So far, trying to optimize likelihood. Could try to optimize APPROXIMATION to likelihood…

Summary of Approaches

- Gradient Ascent
- EM-based (many variants)
- Gibbs sampling
  - Multiple imputation
- Gaussian approximation
- Bound-and-Collapse

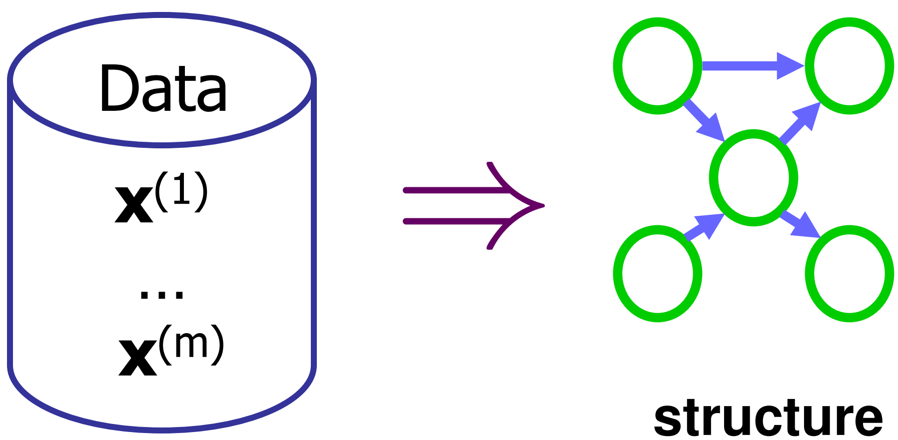
Outline

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

Skip

Learning Bayes Nets

		Structure	
		Known	Unknown
Data	Complete	Easy	NP-hard
	Missing	Hard … EM	Very hard!!



structure

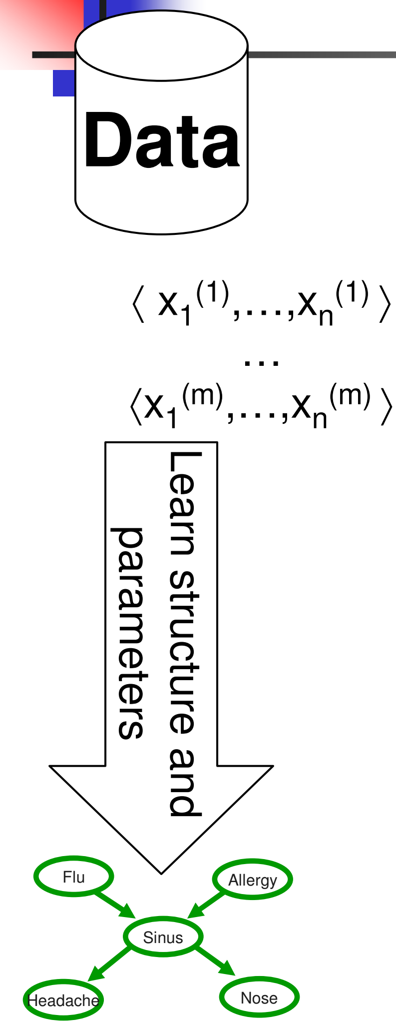
+

CPTs :

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

parameters

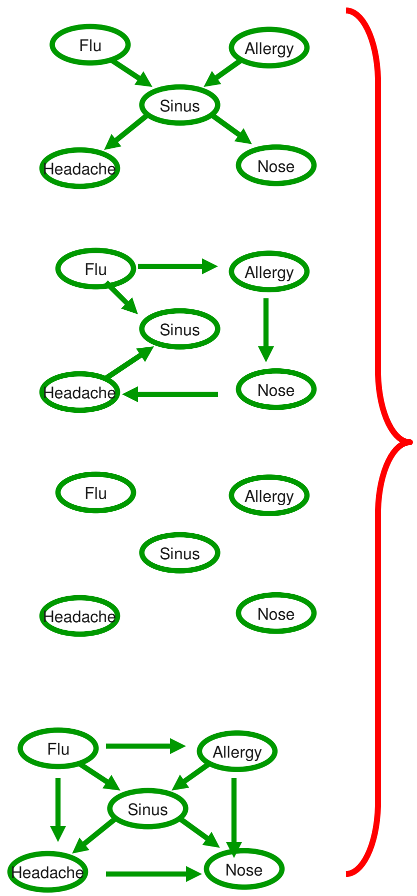
Learning the structure of a BN



- **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.

Score-based Approach

Possible DAG structures (gazillions)



Data

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

$\ldots$

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

Learn Parameters

+

Evaluate ...

Score of each Structure

–15,000

–10,000

–20,000

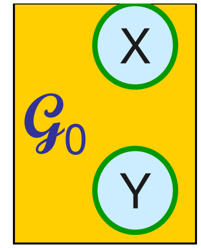
–10,500

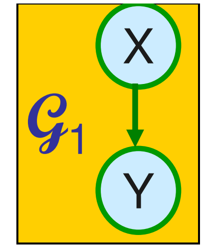
Just use MLE parameters

- $\max_{\mathcal{G},\ \theta_{\mathcal{G}}} L(\langle \mathcal{G}, \theta_{\mathcal{G}} \rangle : \mathcal{S}) =$
  $\max_{\mathcal{G}} \boxed{\max_{\theta_{\mathcal{G}}} L(\langle \mathcal{G}, \theta_{\mathcal{G}} \rangle : \mathcal{S})} =$
  $\max_{\mathcal{G}} L(\langle \mathcal{G}, \boxed{\theta^*_{\mathcal{G}}} \rangle : \mathcal{S})$
- So…
  seek the structure $\mathcal{G}$ that achieves highest likelihood,
  given its MLE parameters $\theta^*_{\mathcal{G}}$

- $\text{Score}(\mathcal{G}, \mathcal{S}) = \log L(\langle \mathcal{G}, \theta^*_{\mathcal{G}} \rangle : \mathcal{S})$

Comparing Models





- $\mathcal{D} = \{\langle x[1],y[1]\rangle, \ldots, \langle x[M],y[M]\rangle\}$

- $\text{Score}(\mathcal{G}_0, \mathcal{S}) = \sum_m \log \theta^*_{x[m]} + \log \theta^*_{y[m]}$
- $\text{Score}(\mathcal{G}_1, \mathcal{S}) = \sum_m \log \theta^*_{x[m]} + \log \theta^*_{y[m] \mid x[m]}$

- $\text{Score}(\mathcal{G}_1, \mathcal{S}) - \text{Score}(\mathcal{G}_0, \mathcal{S})$

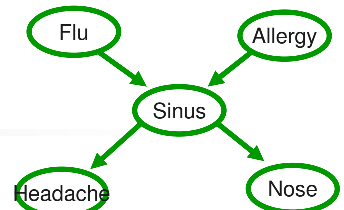
  $= \sum_{x,y} M[x,y] \log \theta^*_{y[m]} - \sum_y M[y] \log \theta^*_{y[m]}$

  $= M \sum_{x,y} p^*(x,y) \log[\, p^*(y|x) / p(y) \,]$

  $= M \; I_{p^*}(X,Y)$

- $I_{p^*}(X,Y)$ = 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 $\mathcal{D}$:

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

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} \log P\left(X_i = x_i^{(j)} \mid \mathbf{Pa}_{X_i} = \mathbf{x}^{(j)}\left[\mathbf{Pa}_{X_i}\right]\right)$$

N_{ijk} θ_{ijk}

$$= \sum_{i=1}^{n} \sum_{x_i, \mathbf{u}} \#(X_i = x_i, \mathbf{Pa}_{X_i} = u) \log P\left(X_i = x_i \mid \mathbf{Pa}_{X_i} = \mathbf{u}\right)$$

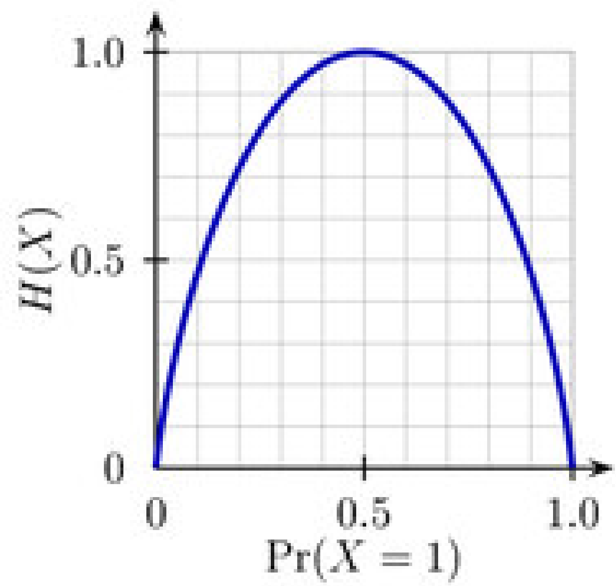
$$= m \sum_{i=1}^{n} \sum_{x_i, \mathbf{u}} \frac{\#(X_i = x_i, \ \mathbf{Pa}_{X_i} = u)}{m} \log P\left(X_i = x_i \mid \mathbf{Pa}_{X_i} = \mathbf{u}\right)$$

$$\boxed{\hat{P}(X_i = x_i, \mathbf{Pa}_{X_i} = u)}$$

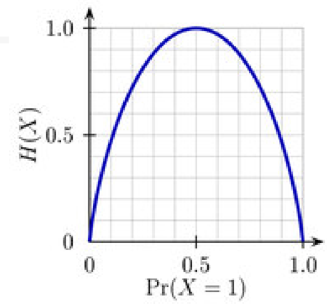
$$= m \sum_{i=1}^{n} \sum_{x_i, \mathbf{u}} \hat{P}(X_i = x_i, \ \mathbf{Pa}_{X_i} = \mathbf{u}) \log P\left(X_i = x_i \mid \mathbf{Pa}_{X_i} = \mathbf{u}\right)$$

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



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

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

Information-theoretic interpretation of maximum likelihood … 2

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

$$\log \hat{P}(\mathcal{D} \mid \theta, \mathcal{G}) = m \sum_i \sum_{x_i, \mathbf{u}} \hat{P}(x_i, \mathbf{Pa}_{x_i,\mathcal{G}} = \mathbf{u}) \log \hat{P}(x_i \mid \mathbf{Pa}_{x_i,\mathcal{G}} = \mathbf{u})$$

$$= m \sum_i -\hat{H}(X_i | \mathbf{Pa}_{x_i,\mathcal{G}})$$

$$= -m \sum_i \hat{H}(X_i | \mathbf{Pa}_{x_i,\mathcal{G}})$$

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

when each $H(X_i | Pa_{X_i,\mathcal{G}})$ is SMALL…

…ie, when parents of X_i are very INFORMATIVE about X_i !

Score for Belief Network

- $\mathcal{I}(X, U) = H(X) - H(X \mid U)$

  $\Rightarrow \; H(X \mid Pa_{X,\mathcal{G}}) = H(X) - \mathcal{I}(X, Pa_{X,\mathcal{G}})$

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

- Log data likelihood

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

- So use score: $\sum_i I(X_i, Pa_{Xi, \mathcal{G}})$

Best Tree Structure

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

- Identify tree with set $\mathfrak{F}$ = { Pa(X) }
  - each Pa(X) is {}, or another variable
- Optimal tree, given data, is

  $\text{argmax}_{\mathfrak{F}}\ m \sum_i I(X_i, Pa(X_i)) - m \sum_i H(X_i)$

  $= \text{argmax}_{\mathfrak{F}} \sum_i I(X_i, Pa(X_i))$

  - … as $\sum_i H(X_i)$ does not depend on structure
- So … want parents $\mathfrak{F}$ s.t.
  - tree structure
  - maximizes $\sum_i I(X_i, Pa(X_i))$

Chow-Liu Tree Learning Alg

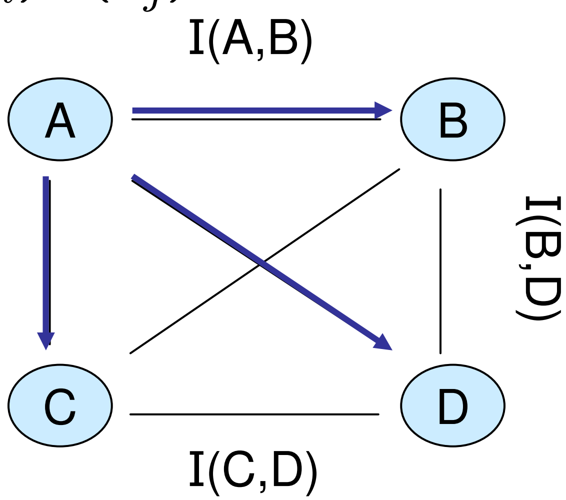
- For each pair of variables X_i, X_j
  - Compute empirical distribution:

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

  - Compute mutual information:

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

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

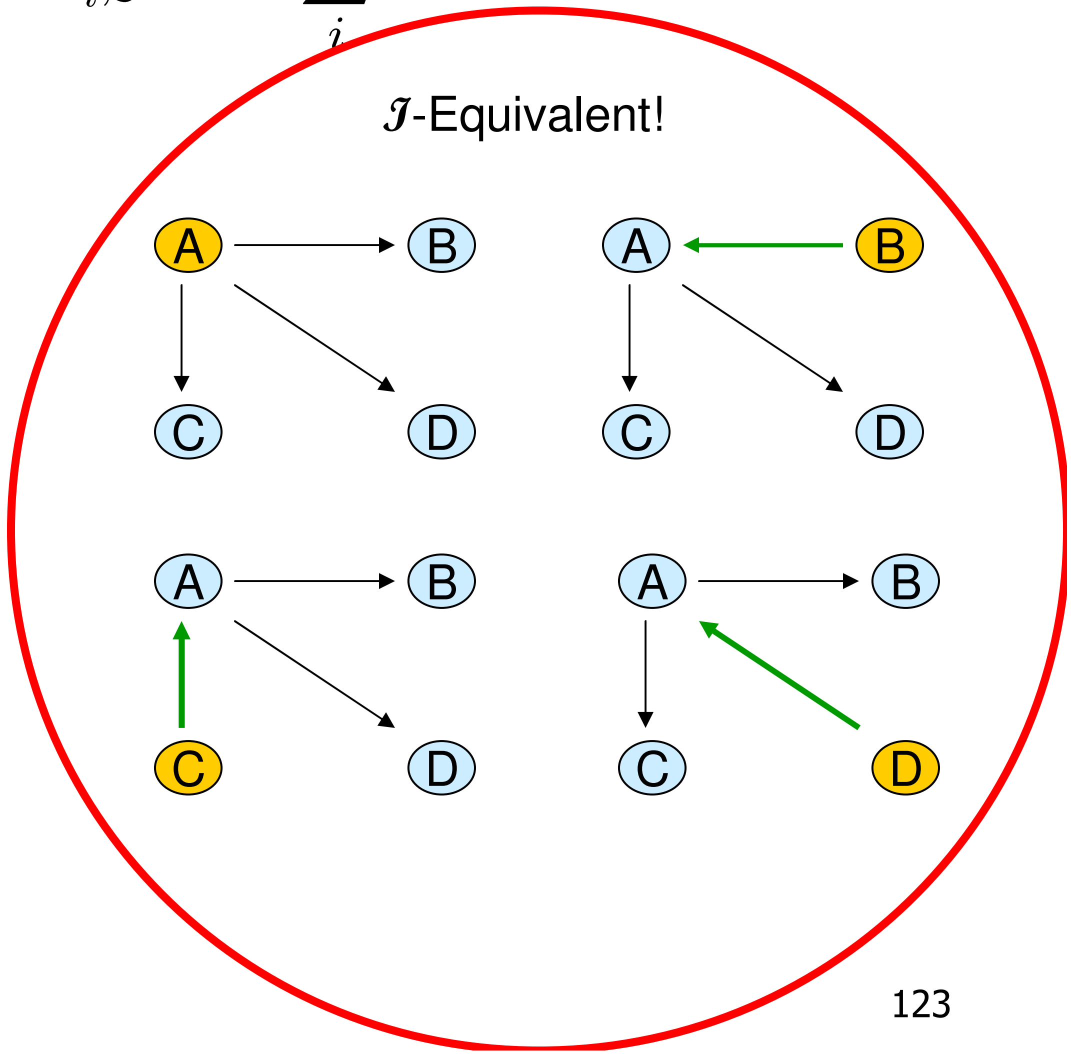


Chow-Liu Tree Learning Alg … 2

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

- 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

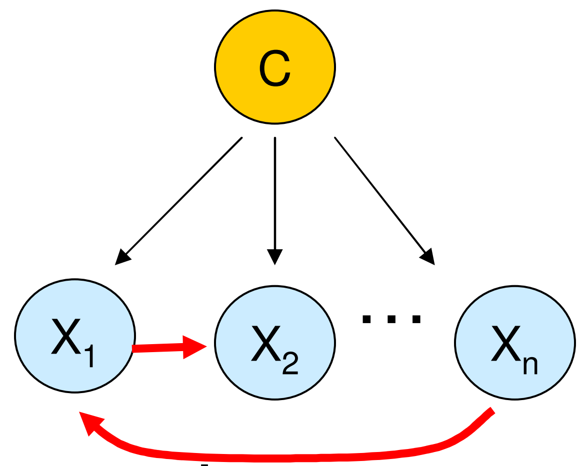


Chow-Liu (CL) Results

- 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 – $\text{argmin}_Q \text{KL}(P; Q)$
- 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

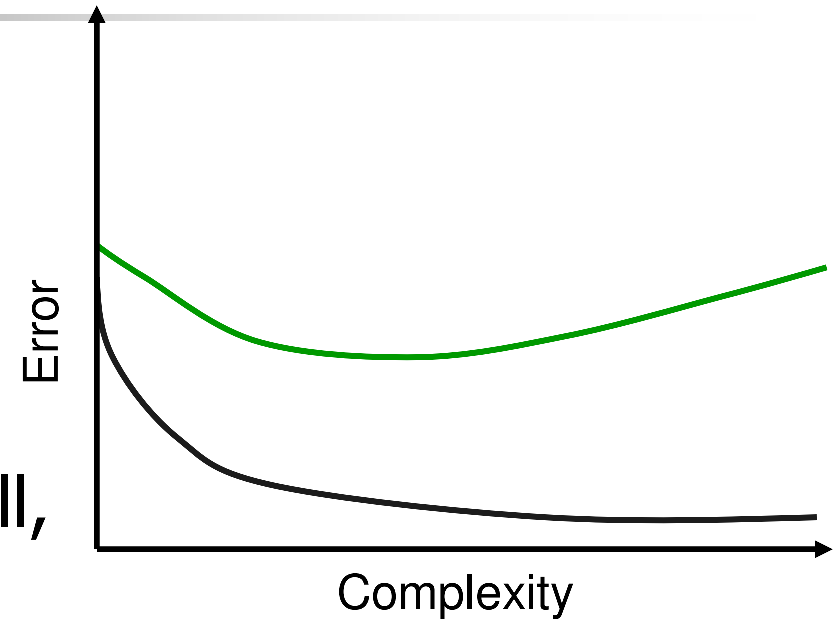
Maximum likelihood score overfits!

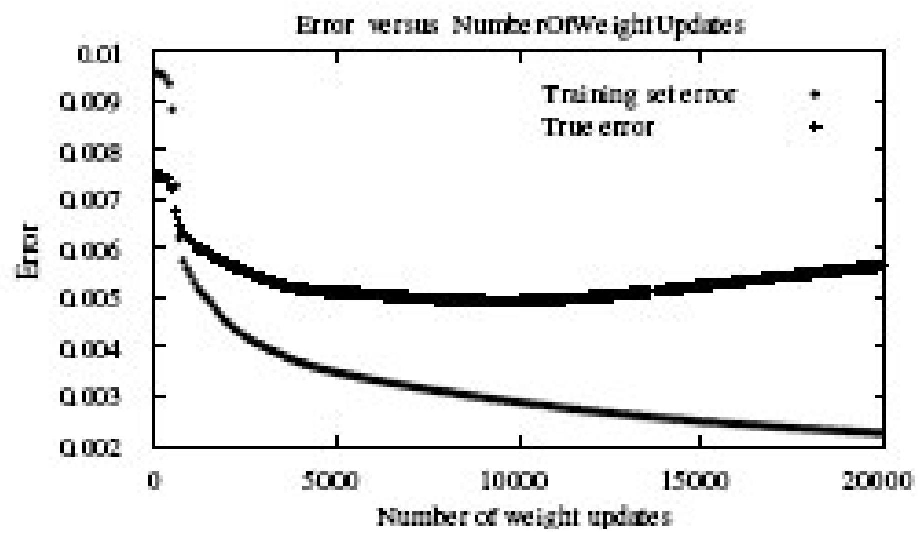
$$\log \hat{P}(\mathcal{D} \mid \theta, \mathcal{G}) = m \sum_i \hat{I}(X_i, \mathbf{Pa}_{X_i,\mathcal{G}}) - m \sum_i \hat{H}(X_i)$$

- Adding a parent never decreases score!!!
  - *Facts:* $H(X \mid Pa_{X,\mathcal{G}}) = H(X) - I(X, Pa_{X,\mathcal{G}})$
    $H(X \mid A) \geq H(X \mid A \cup Y)$
  - $I(X_i, Pa_{Xi,\mathcal{G}} \cup Y) = H(X_i) - H(X_i \mid Pa_{Xi,\mathcal{G}} \cup Y)$
    $\geq H(X_i) - H(X_i \mid Pa_{Xi,\mathcal{G}})$
    $= I(X_i, Pa_{Xi,\mathcal{G}})$
- So score increases as we add edges!
  - Best is COMPLETE Graph
  - … overfit !

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

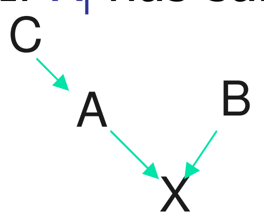
  Goal: Prefer simpler structures... regularization ...
- Posterior over structures given data:
  - $\underbrace{\mathrm{P}(\mathcal{G}|\mathcal{D})}_{\text{Posterior}} \propto \underbrace{\mathrm{P}(\mathcal{D}|\mathcal{G})}_{\text{Likelihood}} \times \underbrace{\mathrm{P}(\mathcal{G})}_{\text{Prior over Graphs}}$
  - $\mathrm{P}(\mathcal{D}|\mathcal{G}) = \int_{\Theta} \mathrm{P}(\mathcal{D} \mid \mathcal{G}, \Theta)\, \underbrace{\mathrm{P}(\Theta|\mathcal{G})}_{\text{Prior over Parameters}}\, d\Theta$

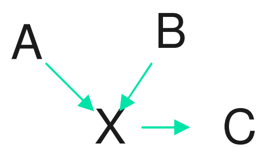
$$\log P(\mathcal{G} \mid D) \approx \log P(\mathcal{G}) + \log \int_{\theta_{\mathcal{G}}} P(D \mid \mathcal{G}, \theta_{\mathcal{G}}) P(\theta_{\mathcal{G}}|\mathcal{G}) d\theta_{\mathcal{G}}$$

Towards a decomposable Bayesian score

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

- **Local and global parameter independence** $\theta_{Y|+x} \perp \theta_X$
- Prior satisfies **parameter modularity**:
  - If X_i has same parents in G and 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!
  - $\log P(\mathcal{G}|\mathcal{D}) = \sum_X \text{ScoreFam}(X \mid Pa_X : \mathcal{D})$

Marginal Probability of Graph

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

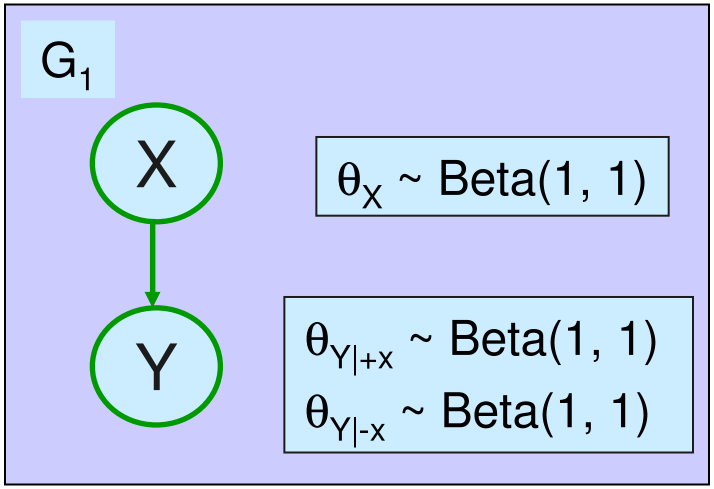
- Given complete data, independent parameters, …

$$\boxed{P(D|G) = \prod_i \prod_{u_i \in Val(Pa_{X_i})} \frac{\Gamma(\alpha^G_{X_i|u_i})}{\Gamma(\alpha^G_{X_i|u_i} + M[u_i])} \prod_{x_i^j \in Val(X_i)} \frac{\Gamma(\alpha^G_{x_i^j|u_i} + M[x_i^j, u_i])}{\Gamma(\alpha^G_{x_i^j|u_i})}}$$

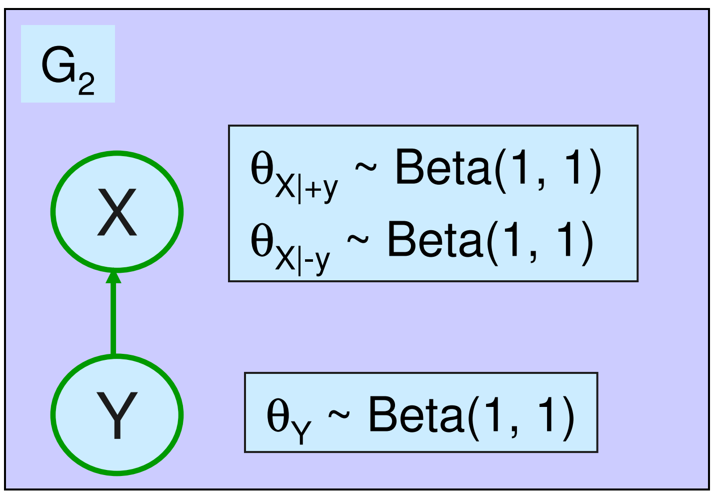
Priors for General Graphs

- For finite datasets, prior is important!
- Prior over structure satisfying prior modularity
  - Eg, $P(\mathcal{G}) \propto c^{|\mathcal{G}|}$ $|\mathcal{G}|$=#edges; $c<1$
- What is good prior over *all* parameters?
  - *K2 prior*: fix $\alpha \in \Re^+$, set $\theta_{X_i|\mathbf{Pa}X_i} \sim \text{Dirichlet}(\alpha, \ldots, \alpha)$
  - 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

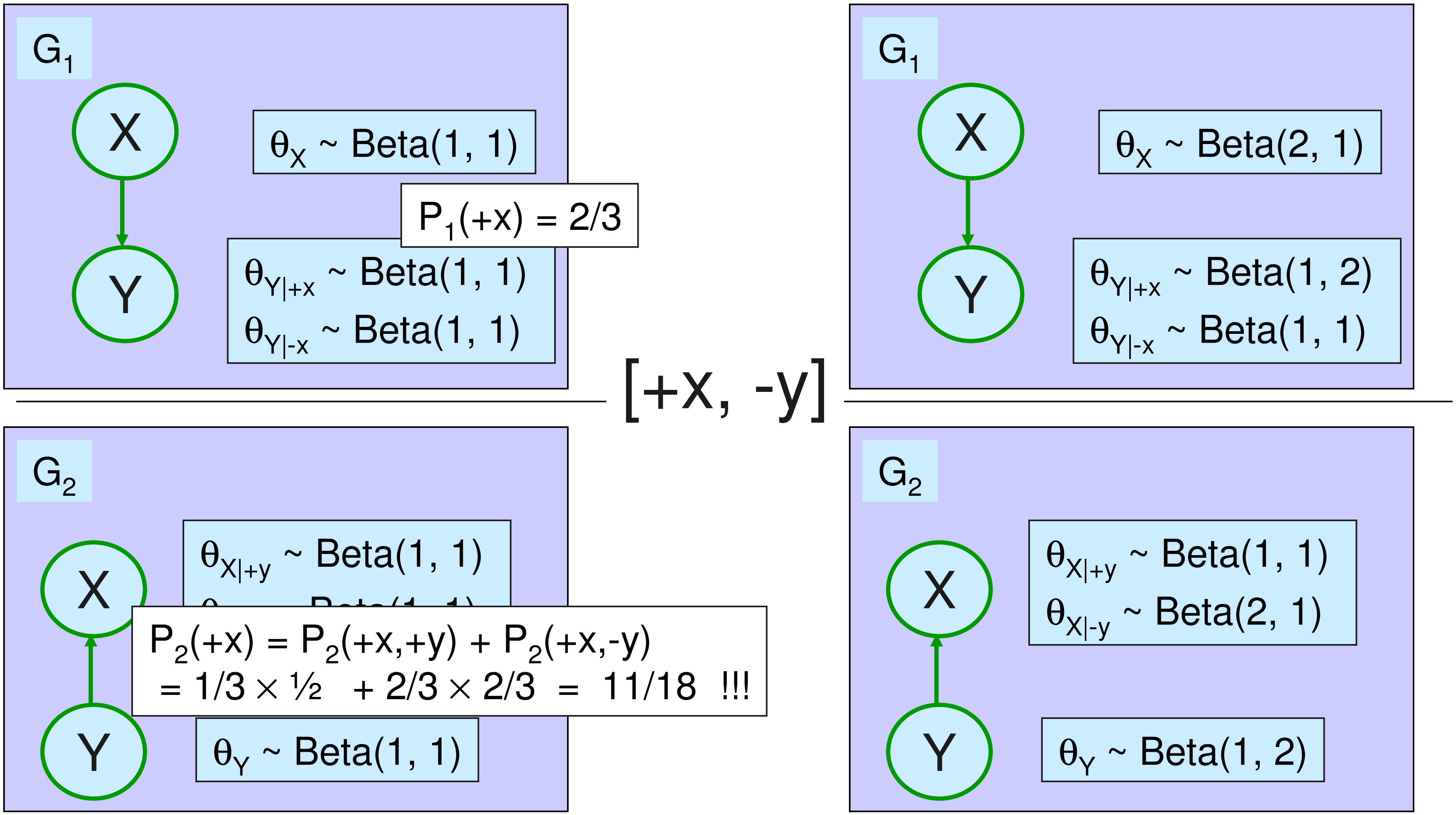


- Does this make sense?
  - EffectiveSampleSize($\theta_{Y|+x}$) = 2
  - But only 1 example ~ "+x" ??

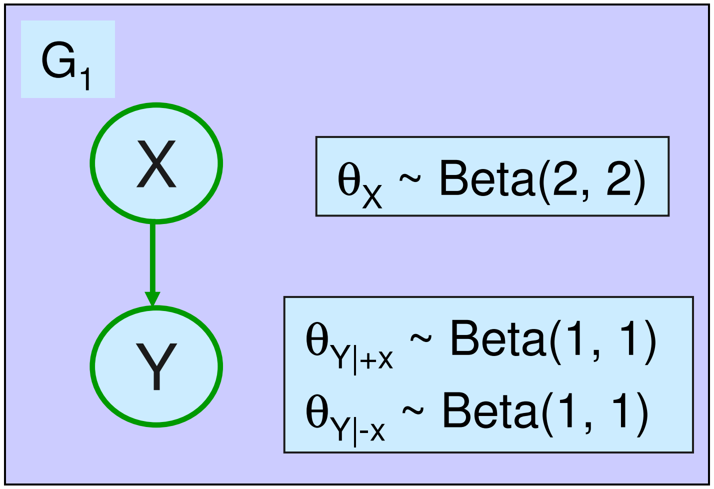


- $\mathcal{I}$-Equivalent structure
- What happens after [+x, -y] ?
  - Should be the same!!

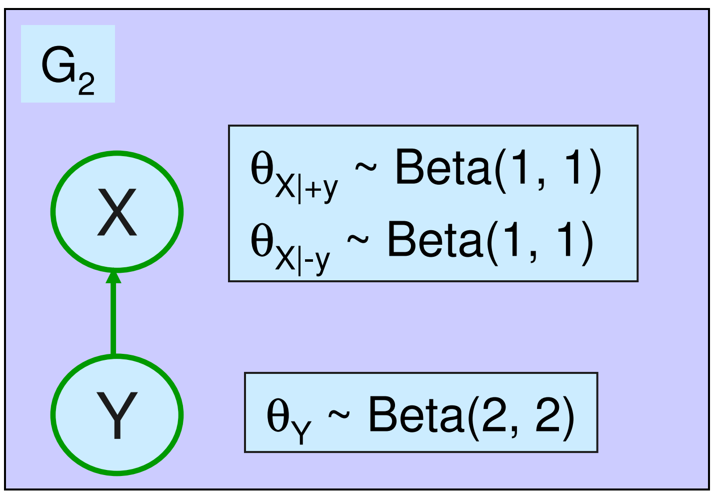
Priors for Parameters



BDe Priors



- This makes more sense:
  - EffectiveSampleSize($\theta_{Y|+x}$) = 2
  - Now ≈∃ 2 examples ~ "+x" ??



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

BDe Priors

G_1

$X \rightarrow Y$

$\theta_X \sim Beta(2, 2)$

$P_1(+x) = 3/5$

$\theta_{Y|+x} \sim Beta(1, 1)$

$\theta_{Y|-x} \sim Beta(1, 1)$

G_1

$X \rightarrow Y$

$\theta_X \sim Beta(3, 2)$

$\theta_{Y|+x} \sim Beta(1, 2)$

$\theta_{Y|-x} \sim Beta(1, 1)$

[+x, -y]

G_2

$Y \rightarrow X$

$\theta_{X|+y} \sim Beta(1, 1)$

$P_2(+x) = P_2(+x,+y) + P_2(+x,-y)$
$= 2/5 \times ½ + 3/5 \times 2/3 = 3/5$!!!

$\theta_Y \sim Beta(2, 2)$

G_2

$Y \rightarrow X$

$\theta_{X|+y} \sim Beta(1, 1)$

$\theta_{X|-y} \sim Beta(2, 1)$

$\theta_Y \sim Beta(2, 3)$

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)
    - $P(X_i, Pa_{X_i}) = P'(x_i) \prod_{x_j \in Pa[X_i]} P'(x_j)$
    - $P(\theta[x_i \mid Pa_{X_i} = u) = Dir(m' P'(x_i{=}1, Pa_{X_i} = u), \ldots, m' P'(x_i{=}k, Pa_{X_i} = u))$
    - Typically, $P'(X_i)$ = 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