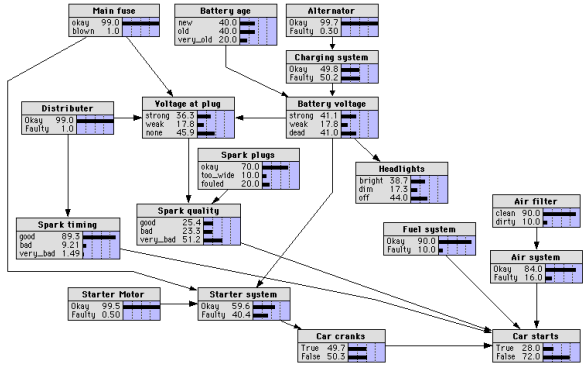
Cmput 466 / 551



Learning Belief Net Parameters

Readings: ≈HTF 17

(Bayesian Networks without the Tears (Charniak))

R Greiner

University of Alberta

Some material taken from C Guesterin (CMU)

Outline

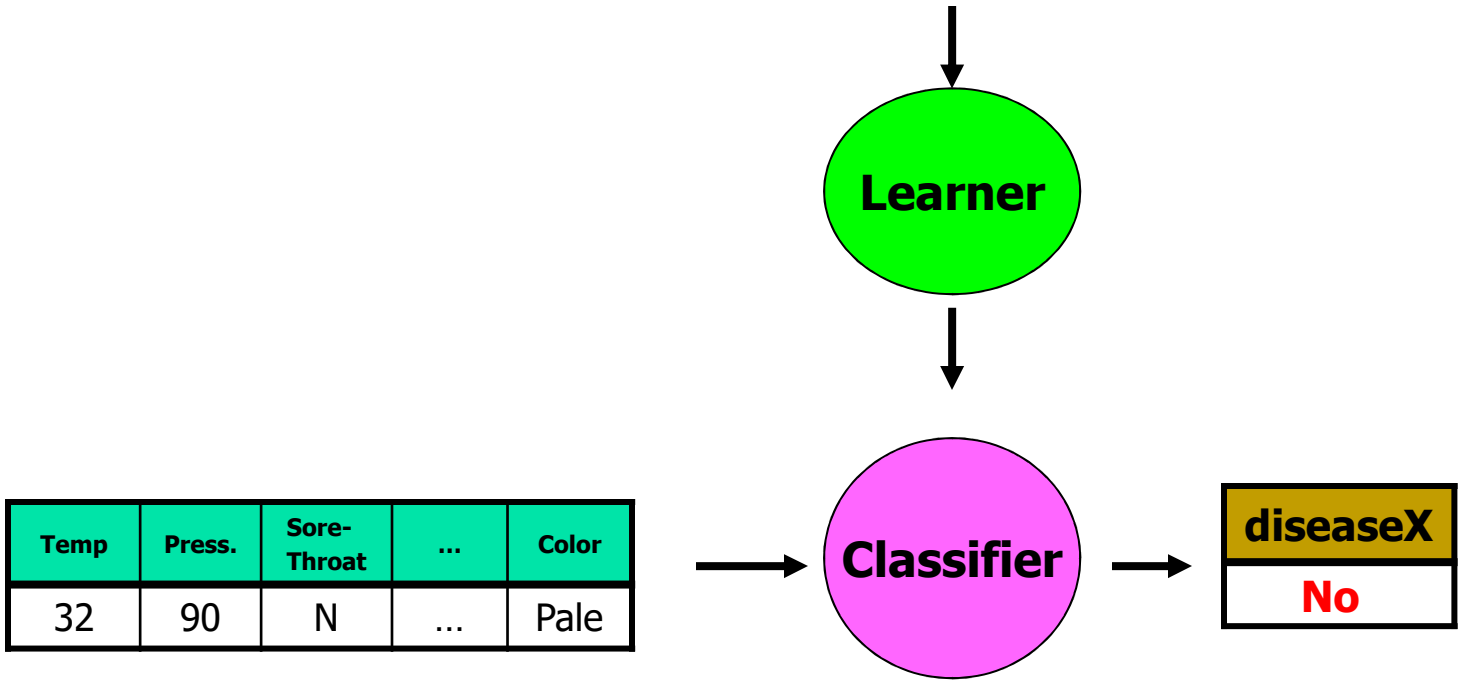
Introduce:
- Density estimation
- KL-divergence … ≈ MLE
- Expectation Maximization
- Gibbs sampling

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

Jump

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

Learning is … Training a Model

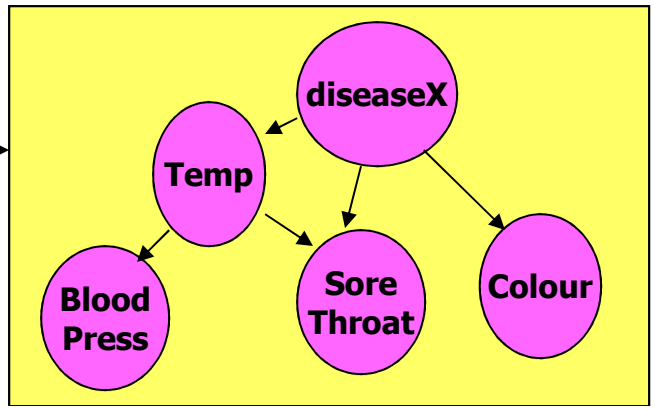
Temp.	Press.	Sore Throat	…	Colour	diseaseX
35	95	Y	…	Pale	No
22	110	N	…	Clear	Yes
:	:			:	:
10	87	N	…	Pale	No

Learner

Then conditionalize, marginalize to answer *any question*:

P(+d | temp=30, BP= 100, …)

Temp	Blood Press.	Sore-Throat	…	Colour	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

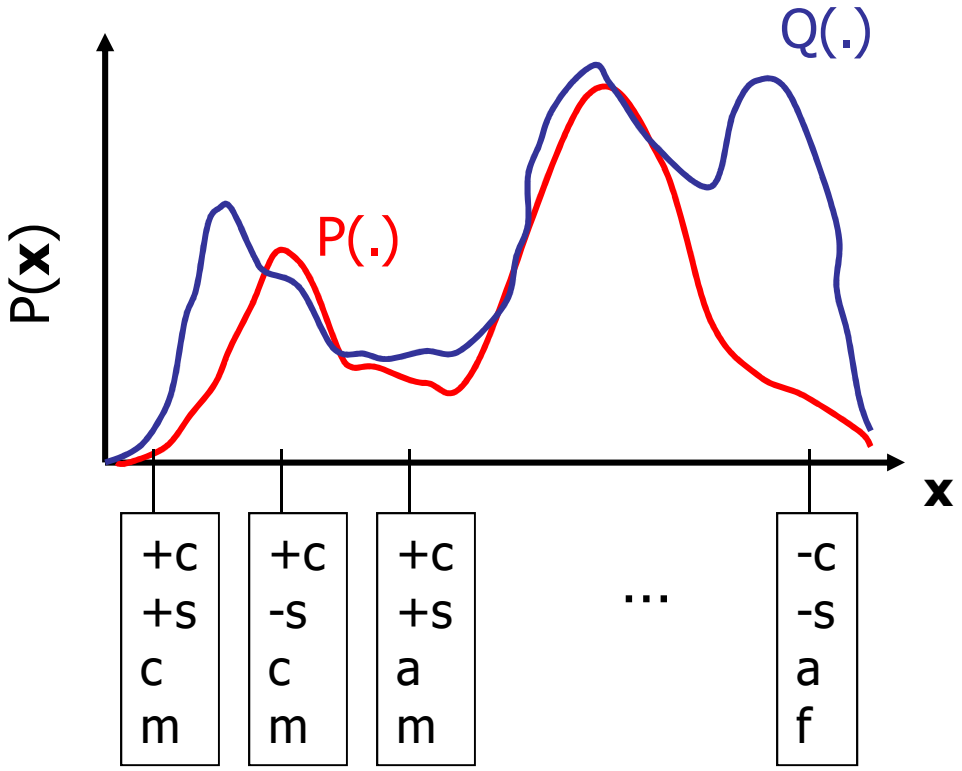
Why Learn Belief Nets?

- 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, …
  - Note: A completely accurate model will produce correct answers to EVERY P(X | Y) query

"Density Estimation"

Generative Learning

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



C	S	A	⋯	G	P	Q
+	y	y	⋯	m	0.3	0.2
+	y	y	⋯	f	0.1	0.15
⋮	⋮	⋮		⋮	⋮	⋮
–	y	o	⋯	f	0.01	0.02
⋮	⋮	⋮		⋮	⋮	⋮

Note: no "y" vs "x" …

KL-Divergence ... ≈ MaxLikelihood

- Seek the BN that minimizes KL-divergence

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

- 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 $\mathcal{D}$

$$BN^* = \text{argmin}_{BN}\ K(\ \mathcal{D};\ BN\ )$$

$$= \text{argmax}_{BN} \sum_x P_{\mathcal{D}}(x) \ln P_{BN}(x)$$

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

$$\approx \text{argmax}_{BN} \sum_{d \in S} \frac{1}{|S|} \ln P_{BN}(\boldsymbol{d})$$

as S drawn from D

$$= \text{argmax}_{BN} \prod_{d \in S} P_{BN}(\boldsymbol{d})$$

$$= \text{argmax}_{BN}\ P_{BN}(\boldsymbol{S})$$

Best Distribution

- If goal is

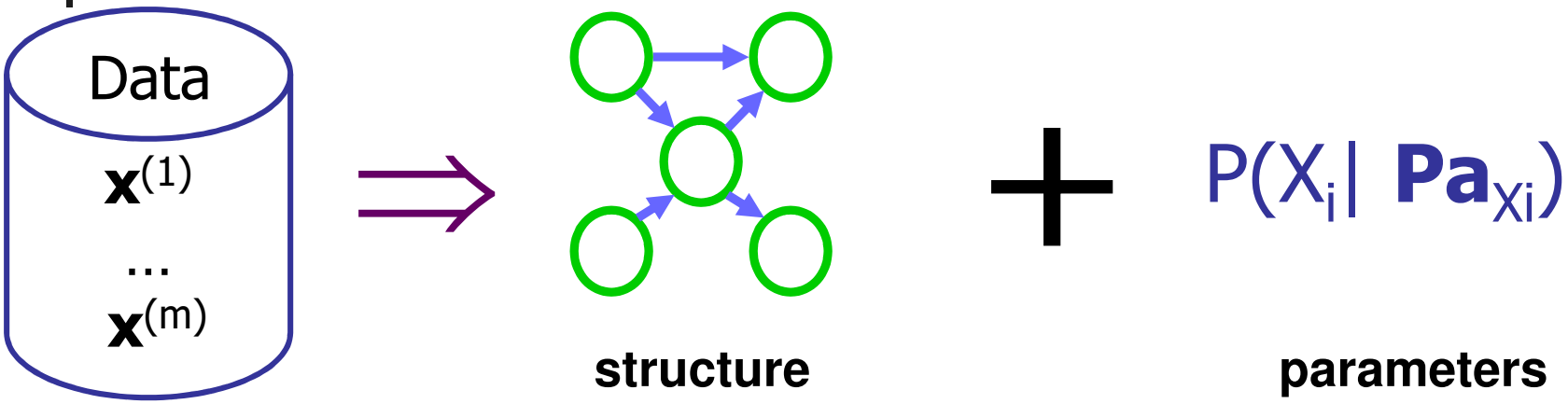
  BN that approximates $\mathcal{D}$:

  Find BN* that maximizes likelihood of data S

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

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

Learning Belief Nets



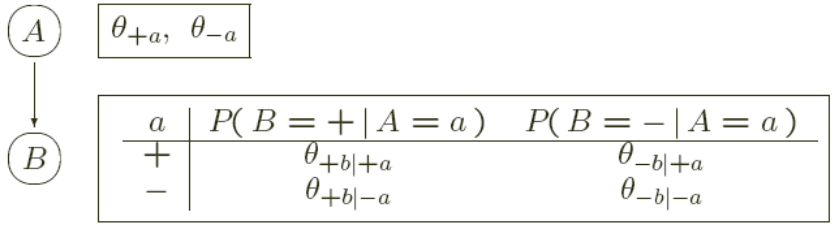
Data	Structure: Known	Structure: Unknown
Complete	**Easy**	**NP-hard**
Missing	**Hard … EM**	**Very hard!!**

Typical (Benign) Assumptions

1. ~~Variables are discrete~~
2. ~~Each case $c_i \in S$ is complete~~
3. Rows of CPtables are independent

   $\theta_A \perp \theta_B$

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

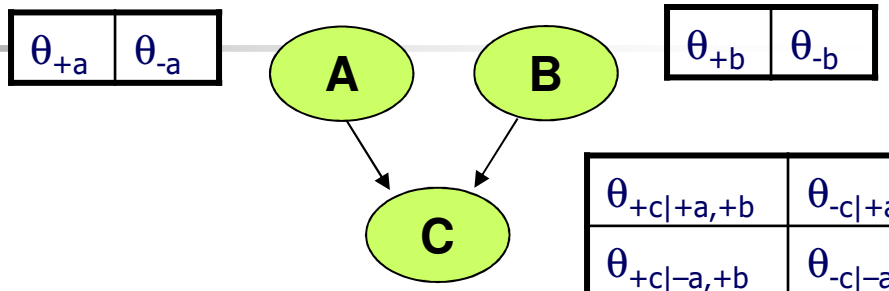


4. ~~Prior $p(\Theta_\chi \mid G)$ is uniform~~
   - $\theta_{B|+a} \sim$ Beta(1,1)

Bayesian Model

- Later: relax Assumptions 1, 2, 4

Learning the CPTs (Frequentist)



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

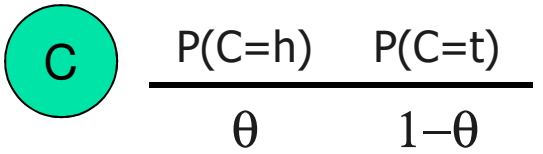
	A	B	C
d_1	1	0	1
d_2	0	1	0
d_3	0	0	1
d_4	1	0	1

WHY????

≈ **REPEAT!!**

One-Node Bayesian Net

- P(Heads) = θ, P(Tails) = $1-\theta$



- Flips are i.i.d.:
  - Independent events
  - Identically distributed according to Binomial distribution
- Sequence S of α_H Heads and α_T Tails

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

Maximum Likelihood Estimation

- **Data:** Observed set 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 $\hat{\theta}$ that maximizes the probability of observed data:

$$\hat{\theta} = \arg\max_{\theta} P(S \mid \theta)$$

$$= \arg\max_{\theta} ln\,[P(S \mid \theta)]$$

Simple "Learning" Algorithm

$$\hat{\theta} = \arg\max_{\theta} ln\,[\,P(\,S\,|\,\theta\,)\,]$$

$$= \arg\max_{\theta} ln\,[\,\theta^{h}\,(1-\theta)^{t}\,]$$

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

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

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

So just average!!!

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

Factoids…

- Recall that, for a Bayesian Network…
  For a COMPLETE instance, $\mathbf{x} = (x_1, \ldots, x_n)$
  $P(\mathbf{x})$ = product of CPtable values
  (one from each variable)

- $\ln a^b = b \ln a$
- $\ln (a \times b) = \ln a + \ln b$

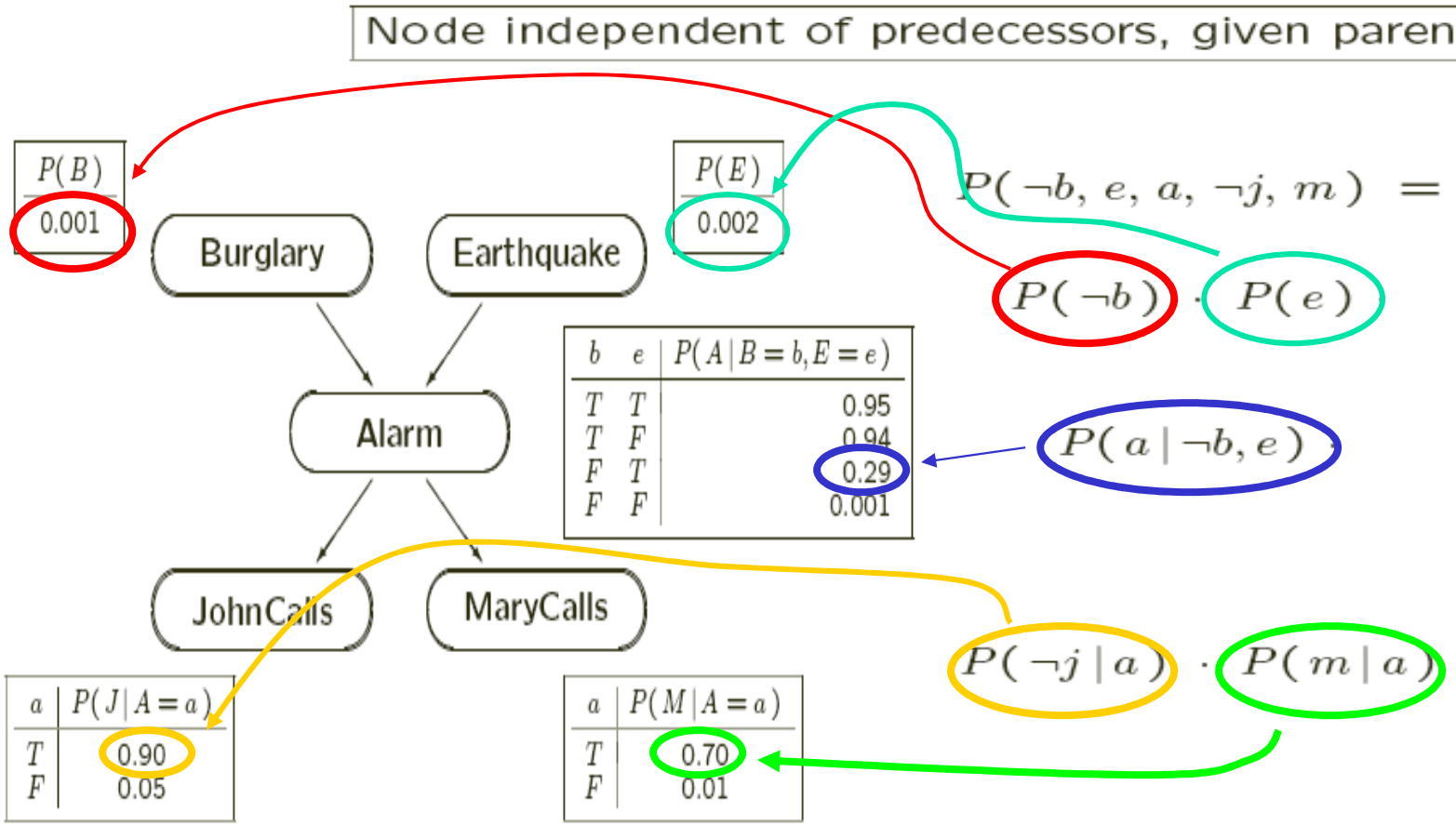
$$\frac{\partial}{\partial \theta} \ln \theta = \frac{1}{\theta}$$

$$\frac{\partial}{\partial \theta} \ln (1-\theta) = \frac{-1}{(1-\theta)}$$

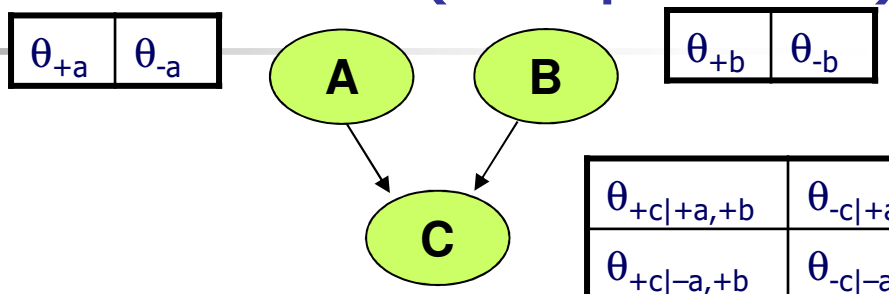
Probability of Complete Instance

$$P(\neg b, e, a, \neg j, m) =$$
$$P(\neg b)\ P(e|\neg b)\ P(a|e,\neg b)\ P(\neg j|a,e,\neg b)\ P(m|\neg j,a,e,\neg b)$$
$$P(\neg b)\ P(e)\ P(a|e,\neg b)\ P(\neg j|a)\ P(m|a)$$
$$0.99 \times 0.02 \times 0.29 \times 0.1 \times 0.70$$

Node independent of predecessors, given parents



Likelihood of the Data (Frequentist)



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

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

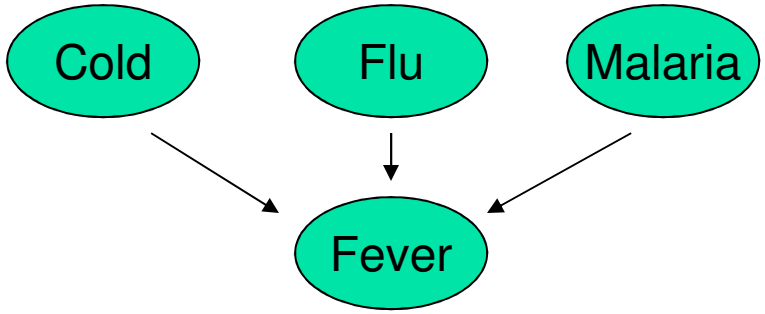
- $P(d_1) = P_\Theta(+a, -b, +c)$
  $= P_\Theta(+a)\ P_\Theta(-b)\ P_\Theta(+c \mid +a, -b)$
  $= \Theta_{+a}\ \Theta_{-b}\ \Theta_{+c|+a,-b}$

- $P(d_2) = P_\Theta(-a, +b, -c)$
  $= P_\Theta(-a)\ P_\Theta(+b)\ P_\Theta(-c \mid -a, +b)$
  $= \Theta_{-a}\ \Theta_{+b}\ \Theta_{-c|-a,+b}$

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

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

Example of Parameter θ_{ijk}



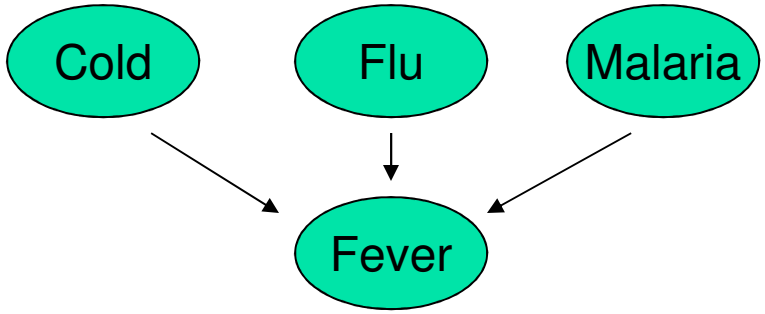
2nd

4th ⇒

Cold	Flu	Malaria	P(Fever=? | Cd, Flu, Mall) 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}	θ_{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]
  - 2nd value of Fever-node – here, "Fever = FALSE"
- Note: $\sum_k \Theta_{ijk} = 1$

Example of Count N_{ijk}



2nd

P(Fever=? | Cd, Flu, Mall)

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}	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}

4th ⇒

- 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

Task#1: Fixed Structure, Complete Tuples

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

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

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

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

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

MLE Values

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

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

- 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

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

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

Algorithm

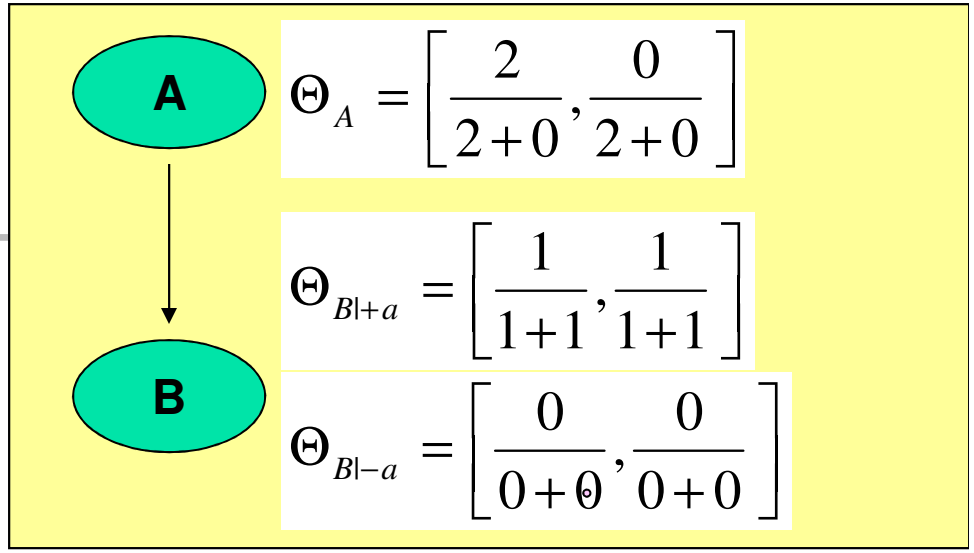
ComputeMLE(graph $\mathcal{G}$, data $\mathcal{S}$):

return MLE parameters $[\theta_{ijk}]$

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

$$\theta_{ijk} = \frac{N_{ijk}}{\sum_r N_{ijr}}$$

Example



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

A	B
+	+
+	−

Huh??

Problems with MLE

Repeat!

- 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…

≈ 50/50 …

Bayesian Learning

Repeat!

$$P(\theta \mid S) \propto P(S \mid \theta)\, P(\theta)$$

posterior *likelihood* *prior*

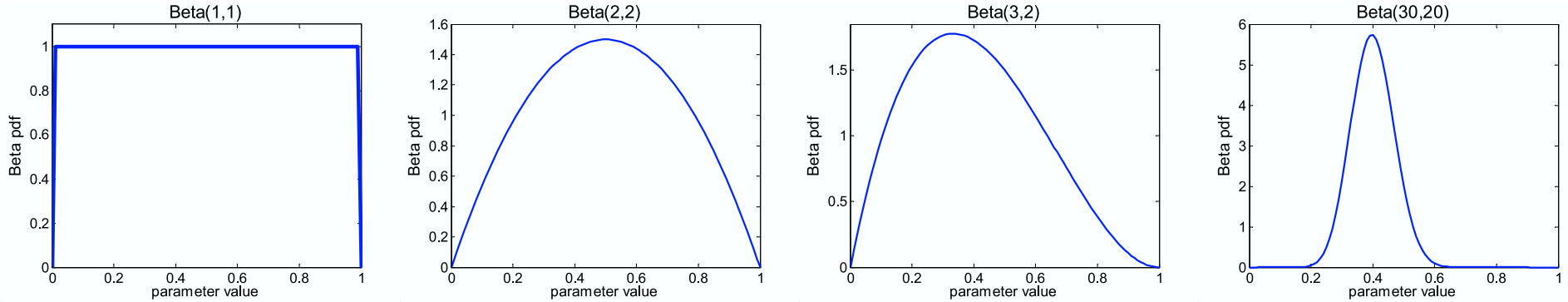
- Likelihood function is simply Binomial:

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

- 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)

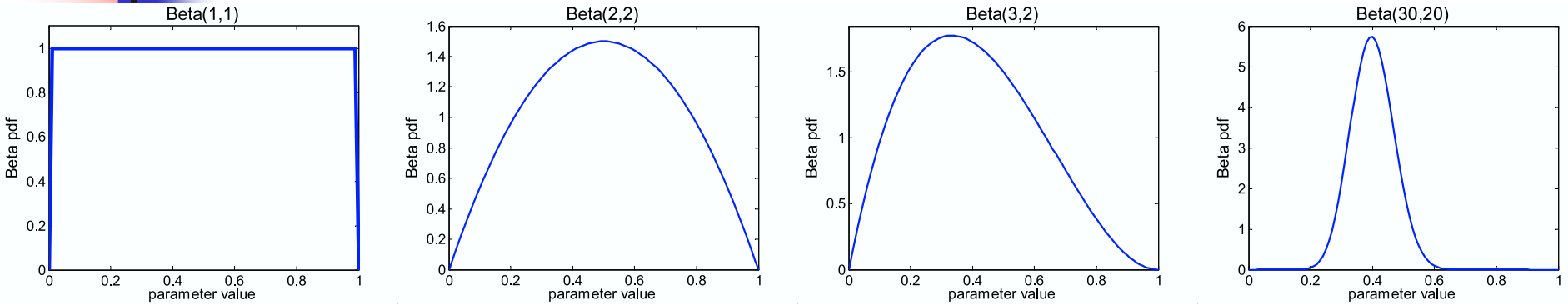
Repeat!



- 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: $\frac{a-1}{a+b-2}$
  - Variance: $\frac{a\,b}{(a+b)^2(a+b-1)}$

Posterior distribution… from Beta

Repeat!



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

Prior P(θ)

Likelihood P(D|θ)

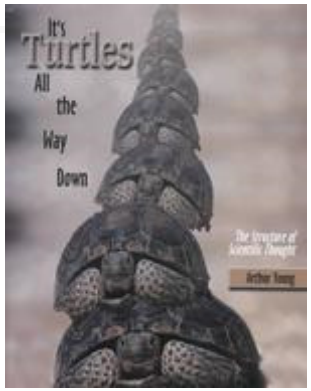
$$= \Theta^{\alpha_H - 1}(1-\Theta)^{\alpha_T - 1} \times \Theta^{m_H}(1-\Theta)^{m_T}$$

$$= \Theta^{\alpha_H + m_H - 1}(1-\Theta)^{\alpha_T + m_T - 1}$$

$$\sim \text{Beta}(\alpha_H + m_H,\ \alpha_T + m_T)$$

Same form! Conjugate!

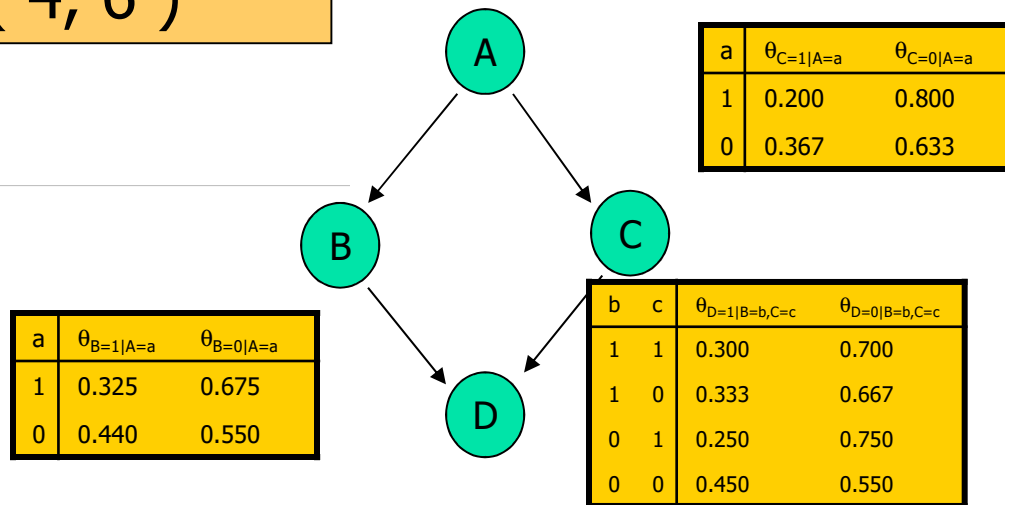
Distribution over Parameter



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

  only a distribution!

$\theta_{A=1}$ ~ Beta(4, 6)

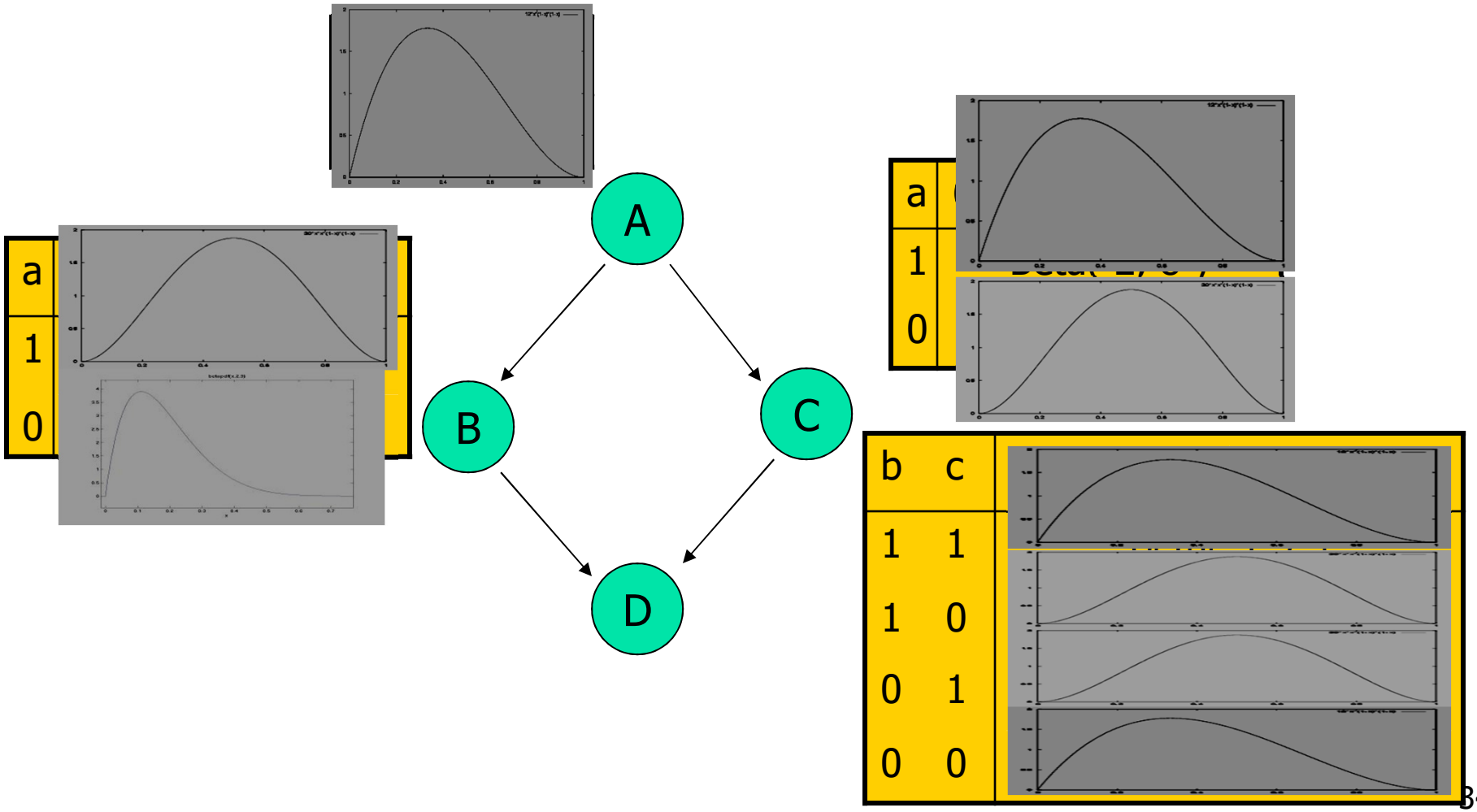


a	$\theta_{C=1\|A=a}$	$\theta_{C=0\|A=a}$
1	0.200	0.800
0	0.367	0.633

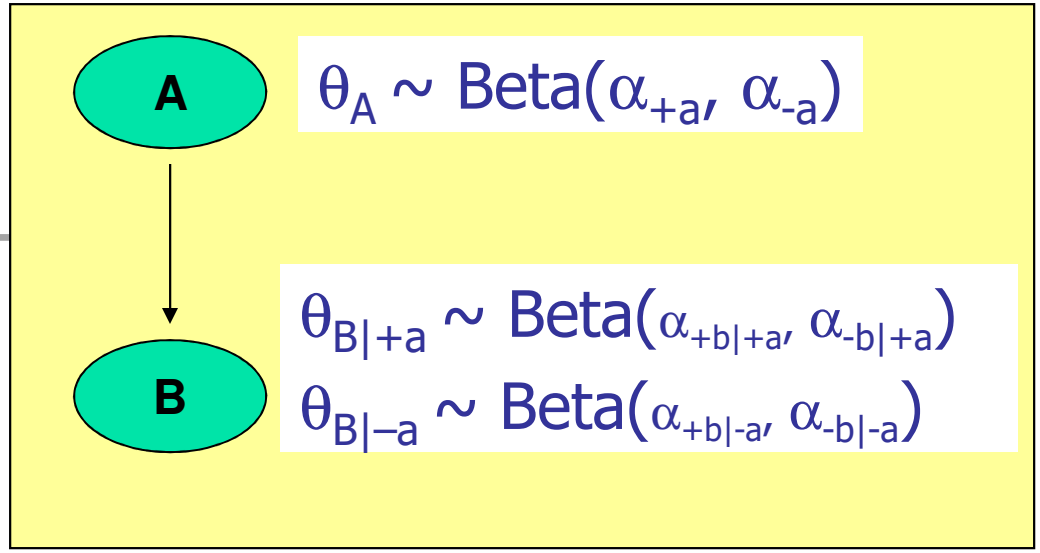
a	$\theta_{B=1\|A=a}$	$\theta_{B=0\|A=a}$
1	0.325	0.675
0	0.440	0.550

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



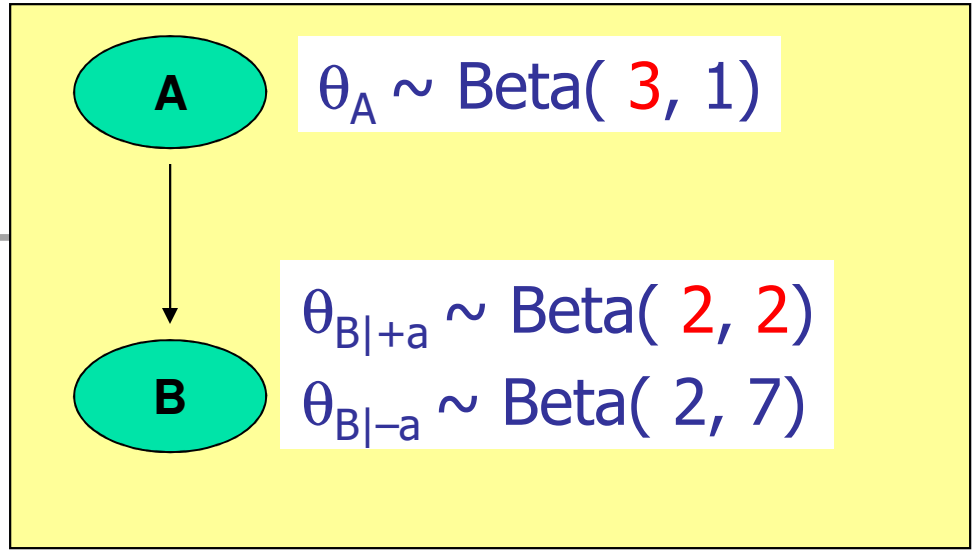
Example



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

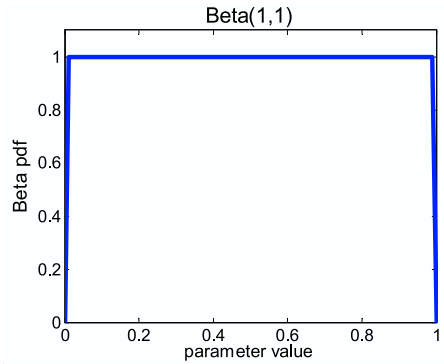
A	B
+	+
+	–

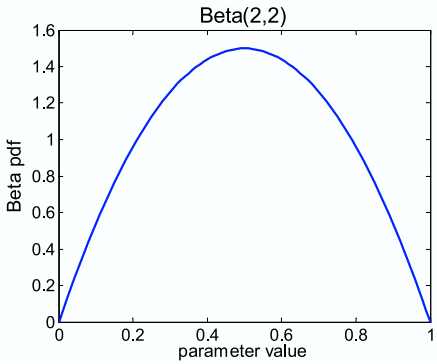
Example



- Buckets
  - $u_{+a} := 1$ (3, 2)
  - $u_{-a} := 1$
  - $u_{+b|+a} := 1$ (2)
  - $u_{-b|+a} := 1$ (2)
  - $u_{+b|-a} := 2$
  - $u_{-b|-a} := 7$

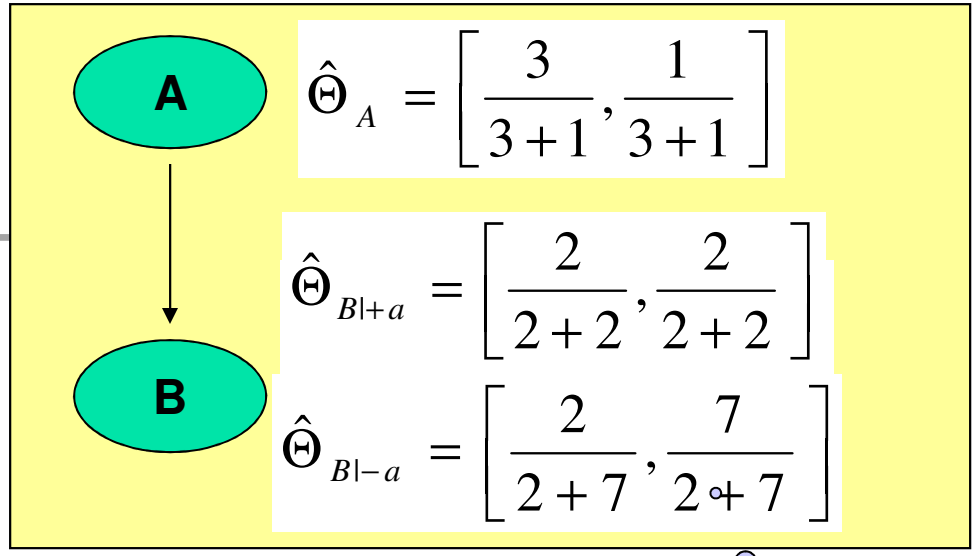
A	B
+	+
+	−





Example

If you want POINT estimates…



- Buckets
  - $u_{+a} := 1$ (3, 2)
  - $u_{-a} := 1$
  - $u_{+b|+a} := 1$ (2)
  - $u_{-b|+a} := 1$ (2)
  - $u_{+b|-a} := 2$
  - $u_{-b|-a} := 7$

A	B
+	+
+	−

Note: no 0/0 issues!

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

Beta Distribution

- Model row-parameter

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

  as *Beta distribution*

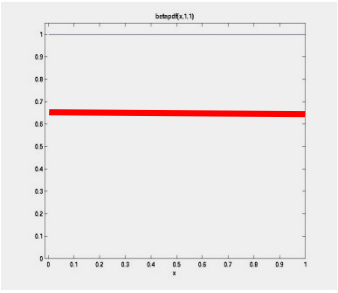
- $\theta_{B|A=1} = \langle \theta_{B=0|A=1}, \theta_{B=1|A=1} \rangle \sim$ Beta(1, 1)

  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)

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

- Now… observe data S :

6 "⟨A=1⟩"

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
⋮	⋮	⋮	⋮

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)

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

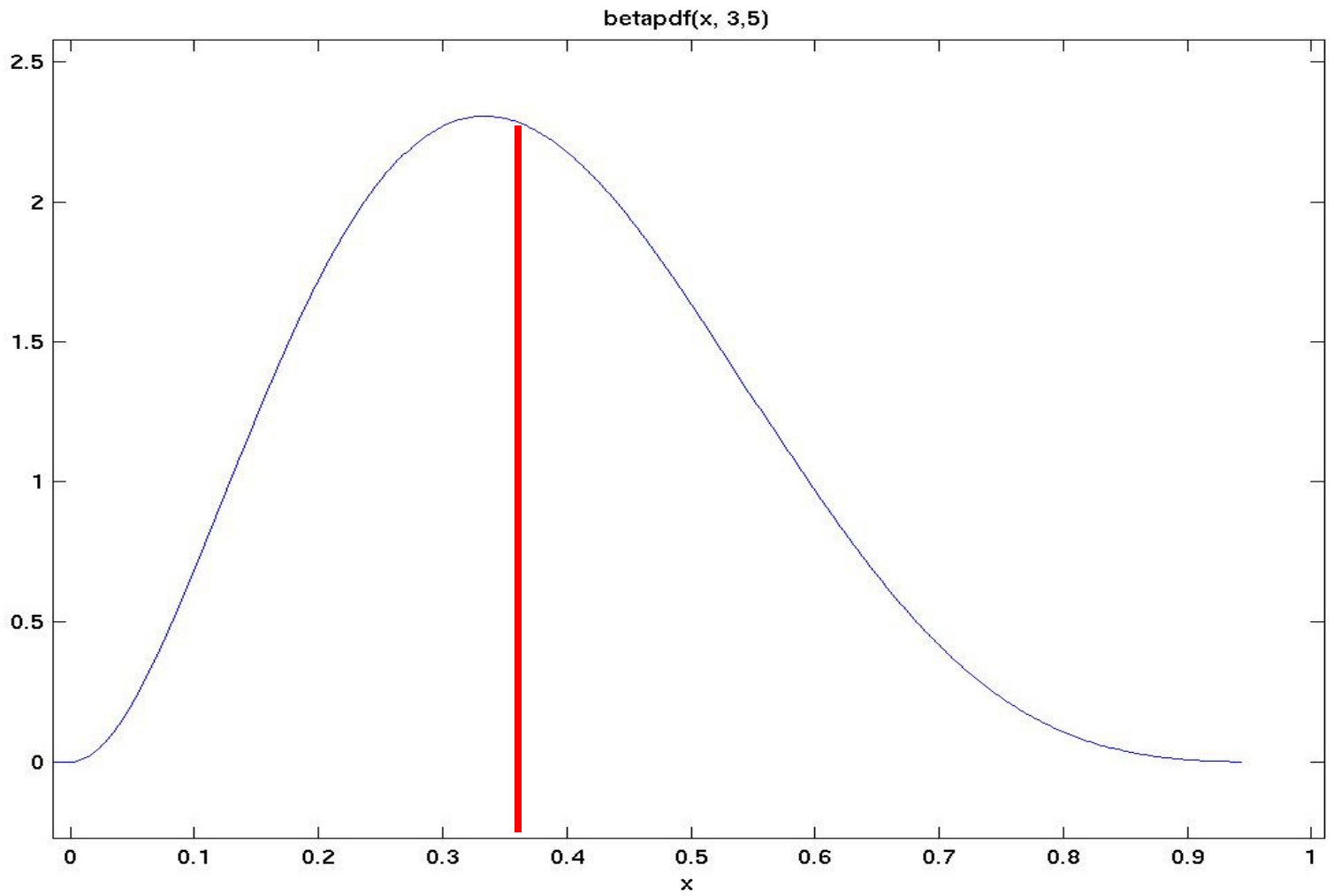
- Then observe data S
  - *2* ⟨A=1, B=1 ⟩
  - *4* ⟨A=1, B=0 ⟩
- *New distribution is*

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

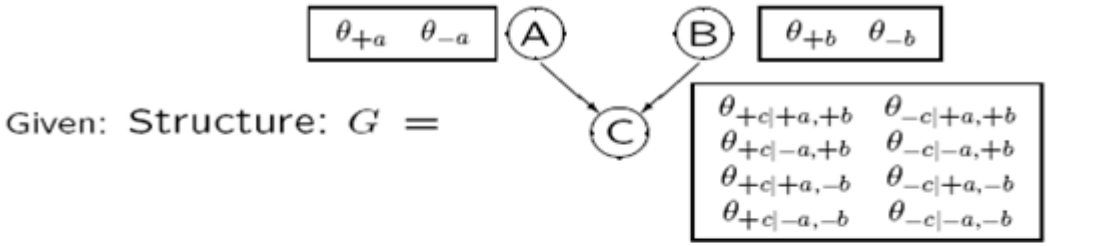
A	B	C	E
1	1	0	1
1	1	1	1
1	0	1	0
1	0	1	0
1	0	0	0
1	0	0	1
0	0	0	1
⋮	⋮	⋮	⋮

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

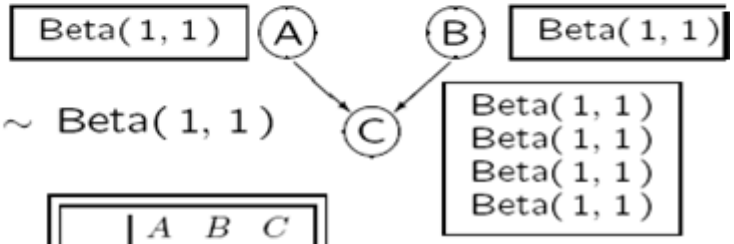
$\theta_{B|A=1} \sim$ Beta(3,5) Distribution



Posterior Distribution of Θ



Given: Structure: G =

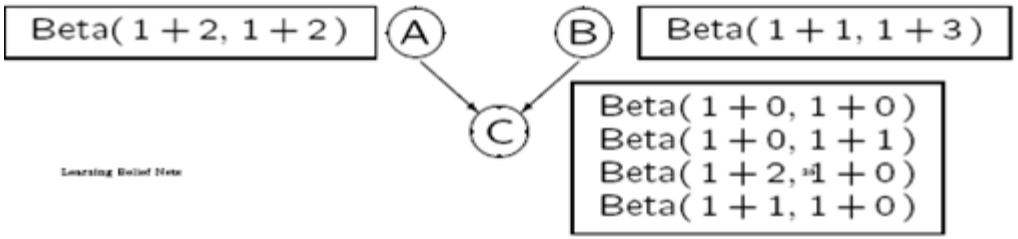


where $\Theta_{X|y} \sim$ Beta(1, 1)

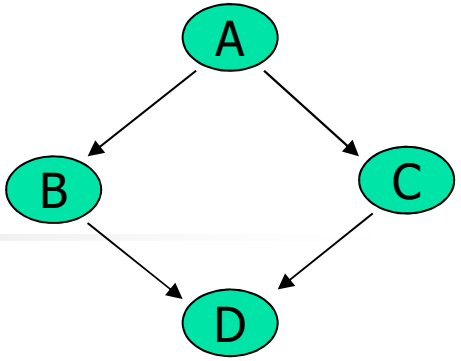
Given sample D =

	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. . .



Posterior Distribution



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

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

- Compare to Frequentist:

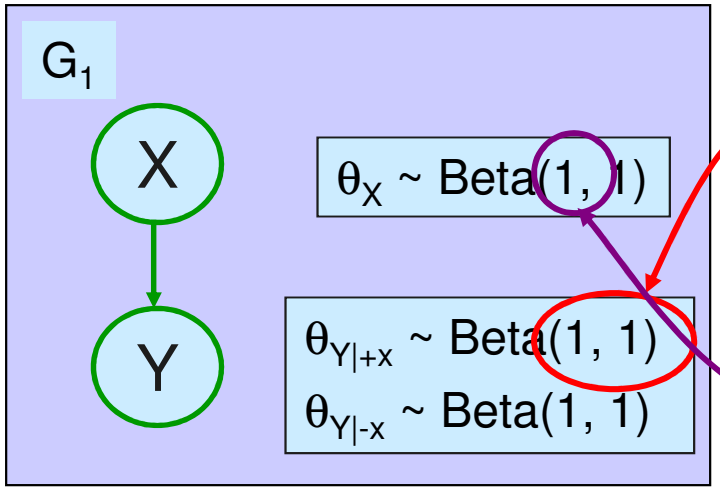
$$\hat{\theta}_{ijk} = \frac{N_{ijk}}{\sum_r N_{ijr}}$$

Algorithm

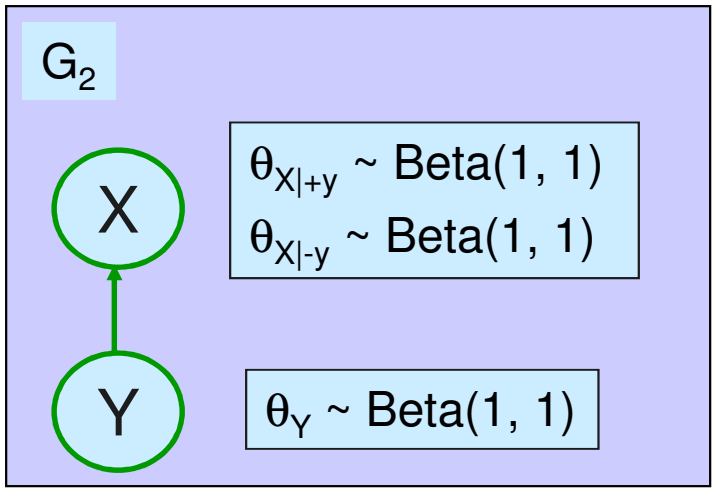
ComputePosterior(graph G, data S, priors $[\alpha_{ijk}]$):
return posterior parameters $[u_{ijk}]$

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

Priors for Parameters

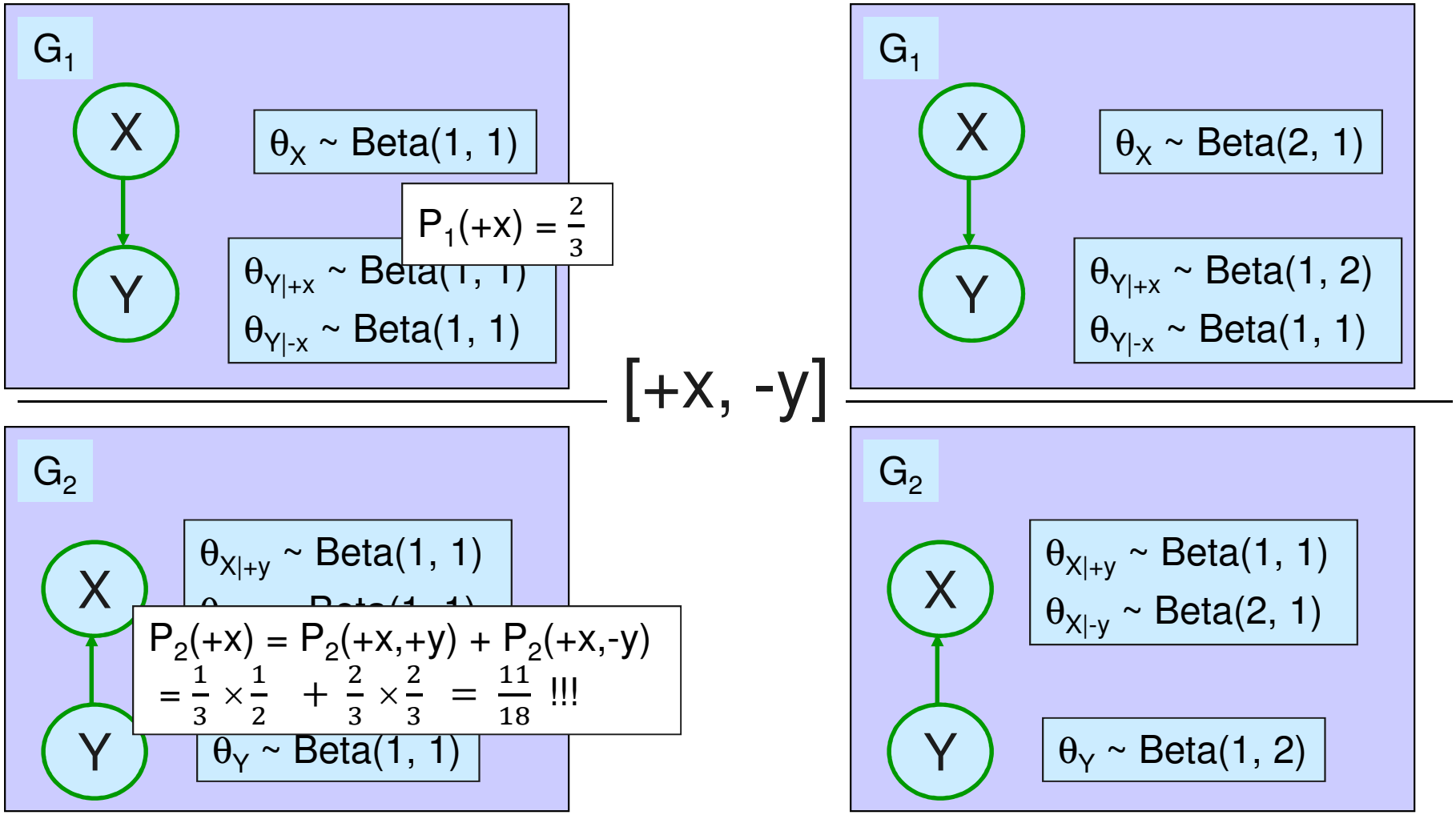


- Does this make sense?
  - EffectiveSampleSize($\theta_{Y|+x}$) = 2
  - But only 1 example ~ "+x" ??

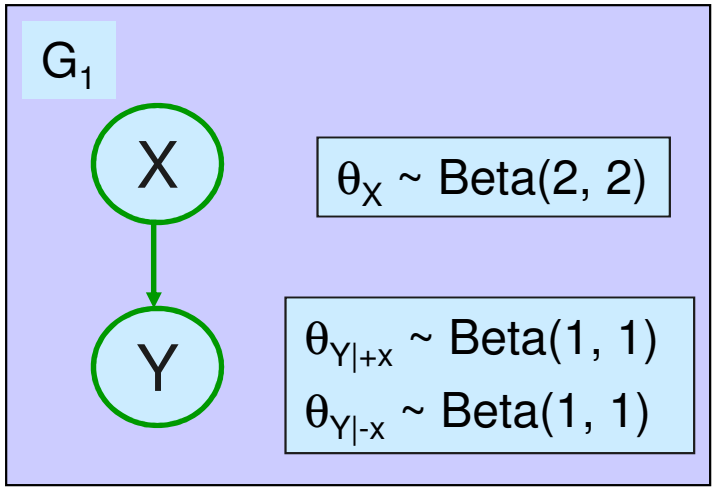


- $\mathcal{I}$-Equivalent structure
- What happens after [+x, -y] ?
  - Should be the same!!

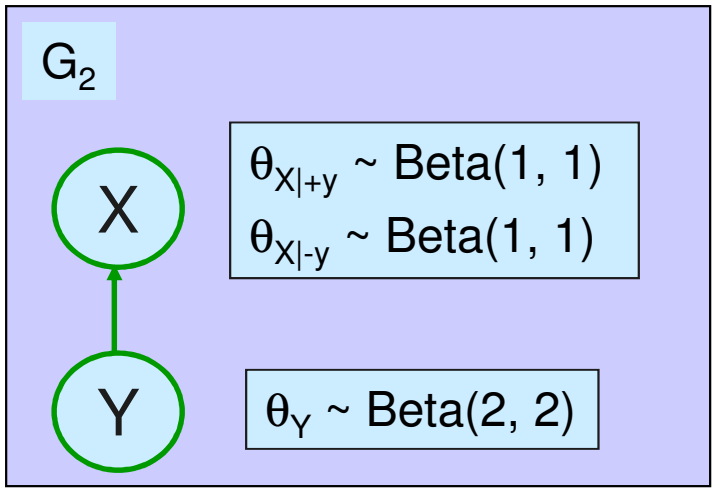
Priors for Parameters



BDe Priors

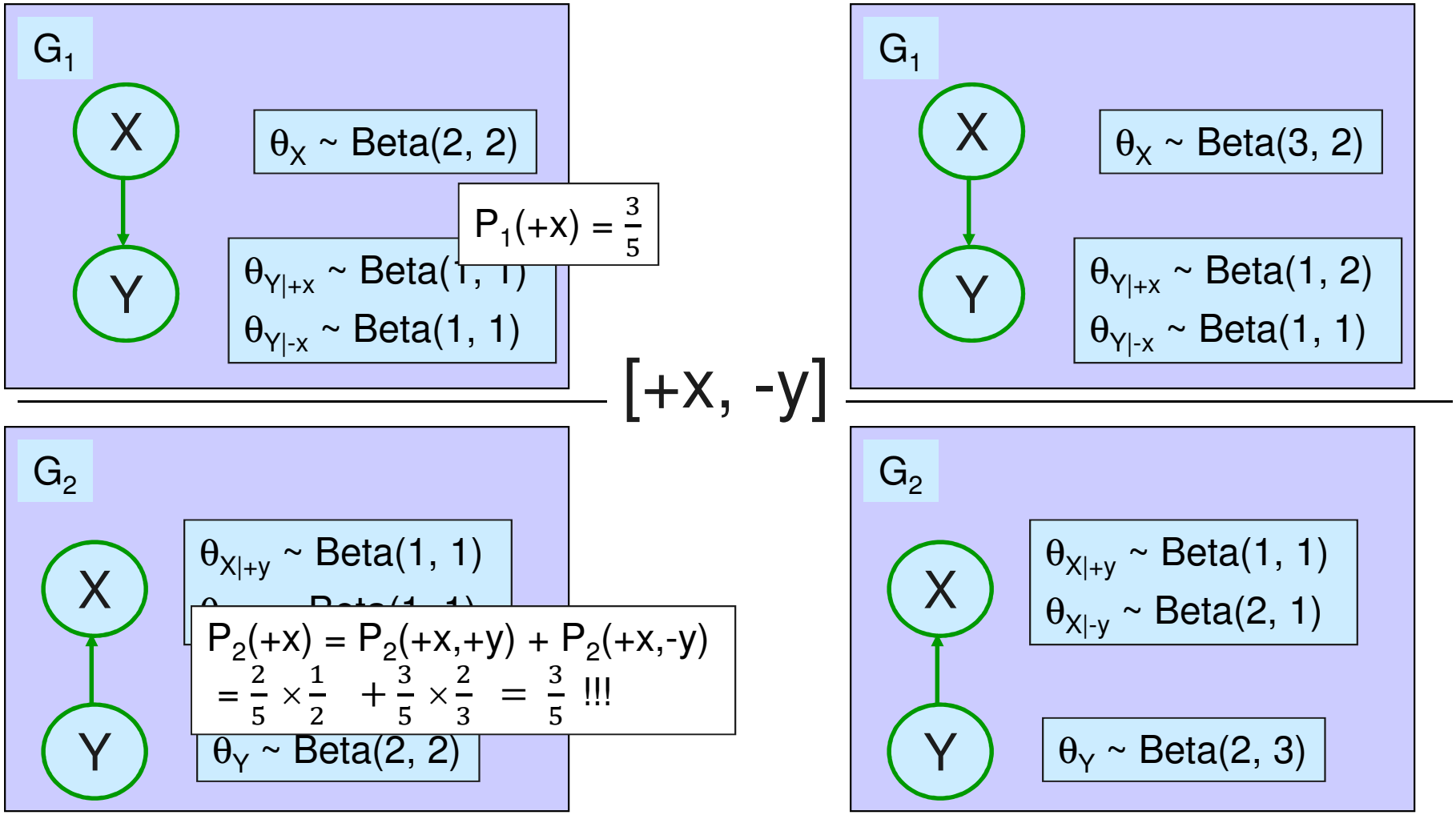


- This makes more sense:
  - EffectiveSampleSize($\theta_{Y|+x}$) = 2
  - Now ≈ 2 examples ~ "+x" ??



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

BDe Priors


G1
X
Y
θX ~ Beta(2, 2)
P1(+x) = 3/5
θY|+x ~ Beta(1, 1)
θY|-x ~ Beta(1, 1)
G1
X
Y
θX ~ Beta(3, 2)
θY|+x ~ Beta(1, 2)
θY|-x ~ Beta(1, 1)
[+x, -y]
G2
X
Y
θX|+y ~ Beta(1, 1)
P2(+x) = P2(+x,+y) + P2(+x,-y)
= 2/5 × 1/2 + 3/5 × 2/3 = 3/5 !!!
θY ~ Beta(2, 2)
G2
X
Y
θX|+y ~ Beta(1, 1)
θX|-y ~ Beta(2, 1)
θY ~ Beta(2, 3)

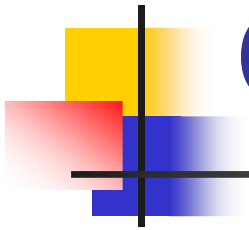
BDe Prior

- View Dirichlet parameters as "fictitious samples" – equivalent sample size
- Pick a fictitious sample size m′
- For each possible family, define a prior distribution $P(X_i, \boldsymbol{Pa}_{X_i})$
  - Represent with a BN
  - Usually independent (product of marginals)
    - $P(X_i, \boldsymbol{Pa}_{X_i}) = P'(x_i) \prod_{x_{j \in \boldsymbol{Pa}[X_i]}} P'(x_j)$
    - $P(\theta[x_i \mid \boldsymbol{Pa}_{X_i} = u) = Dir(m' P'(x_i=1, \boldsymbol{Pa}_{X_i} = u), \ldots, m' P'(x_i=k, \boldsymbol{Pa}_{X_i} = u))$
    - Typically, $P'(X_i)$ = uniform

Summary: Parameter Learning

- MLE:
  - score decomposes according to CPTs
  - optimize each CPT separately
- Bayesian parameter learning:
  - motivation for Bayesian approach
  - Bayesian prediction
  -
  -
- Bayesian learning for BN parameters
  - Global parameter independence
  - BDe if and only if score equivalence
  -
  -
  - 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
    - Gradient Descent
    - EM
    - Gibbs
  - Learning Structure

#2: Known structure, Missing data

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

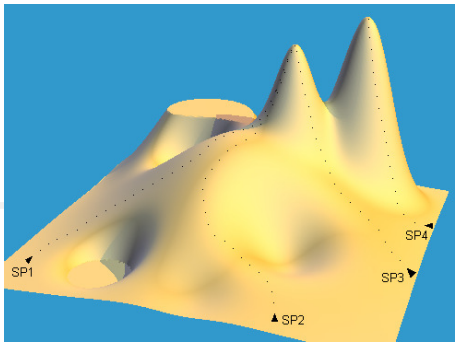
$$\mathcal{S} = \left\{ \begin{array}{lccc} c_1: & \langle * & \cdots & c_{1N} \rangle \\ c_2: & \langle c_{21} & \cdots & * \rangle \\ \vdots & \langle \vdots & c_{ij} & \vdots \rangle \\ c_m: & \langle c_{m1} & \cdots & c_{mN} \rangle \end{array} \right\}$$

incomplete

complete

- What if $\mathcal{S}$ is **incomplete** ?
  - Some $c_{ij} = *$
  - $c_{iK} = * \; \forall$ i (ie, X_K never seen... "Hidden variables")
- 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 expansion…

$$f_{\text{obj}}(\theta) \approx f_{\text{obj}}(\theta^0) + (\theta - \theta^0)\,\nabla f_{\text{obj}}(\theta^0)$$

Need derivative!

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

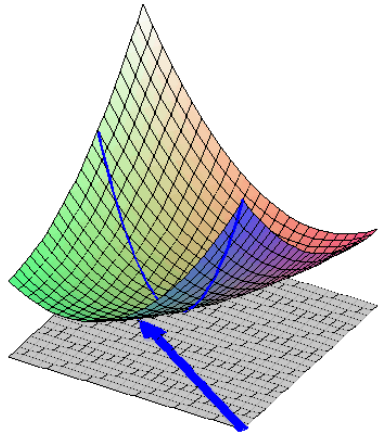
1 $t \leftarrow 1$
2 **do**
3 $\quad \theta^{t+1} \leftarrow \theta^t + \eta \nabla f_{\text{obj}}(\theta^t)$
4 $\quad t \leftarrow t + 1$
5 **while** $\|\theta^t - \theta^{t-1}\| > \delta$
6 **return** (θ^t)

Issues with Gradient Ascent

- Constraints
  - $\Theta_{ijk} \in [0,1]$
  - $\sum_r \Theta_{ijr} = 1$
  - But ... $\Theta_{ijk} \mathrel{+}= \alpha \, \Delta\Theta_{ijk}$ could violate constraints
  - Use $\lambda_{ijk} = \log(\theta_{ijk})$
    - Find best λ_{ijk} ... ignore constraints ...

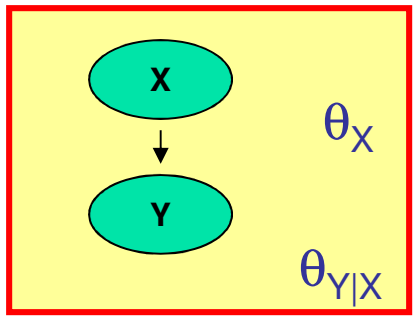
- Lots of Tricks for efficient ascent
  - Line Search
  - Conjugate Gradient
  - ...

  [See earlier notes on optimization]

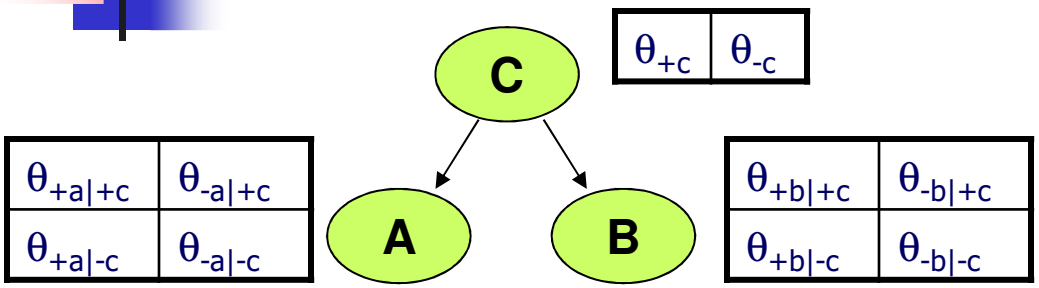


Expectation Maximization (EM)

- EM is designed to find most likely θ, given incomplete data !
- Recall simple Maximization needs counts: $\#(+x, +y)$, … for $N_{+y|+x}$, …
- 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
- Compute "expected sufficient statistics": $E_\theta[N_{ijk}]$



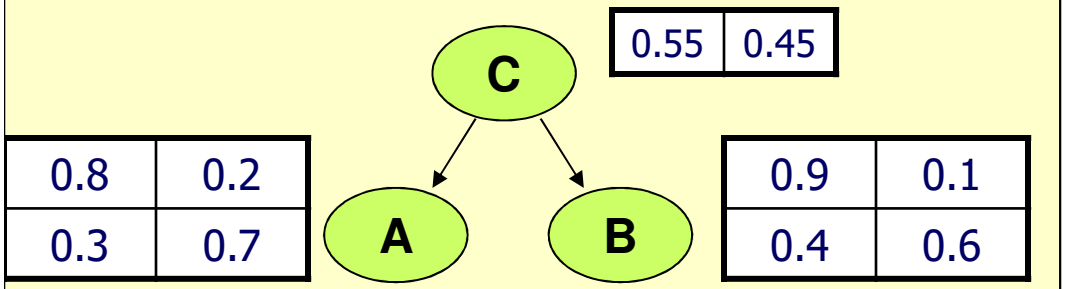
EM Approach – E Step



Sample S =

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

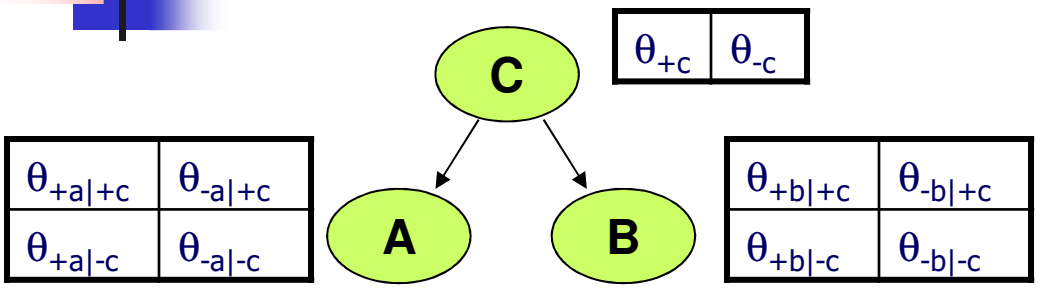
Guess initial values θ^0



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	0.2×0.1
0	1	1	0.2×0.9
1	0	1	0.8×0.1
1	1	1	0.8×0.9

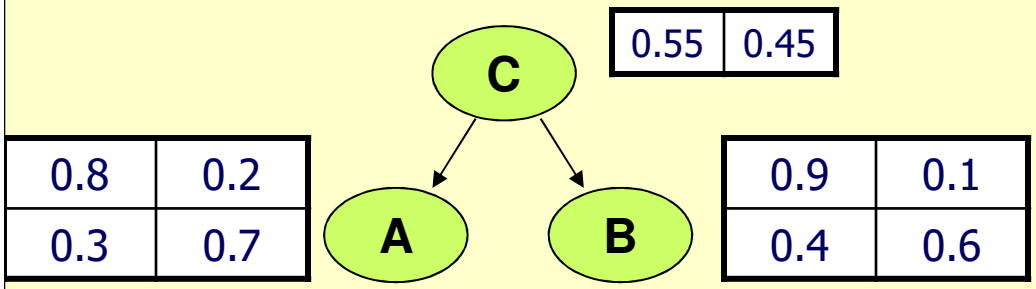
EM Approach – E Step



Sample S =

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

Guess initial values θ^0



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	0.2×0.1
0	1	1	0.2×0.9
1	0	1	0.8×0.1
1	1	1	0.8×0.9

$E_{\theta^0}[\, N_{+b|+c} \,] = 0.9 + (0.2 \times 0.9) + (0.8 \times 0.9)$

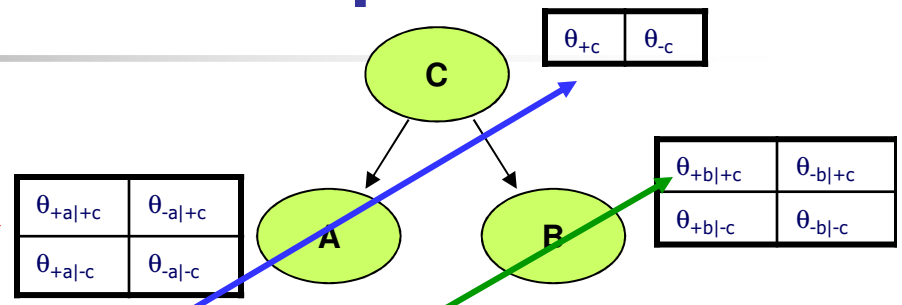
$E_{\theta^0}[\, N_{-b|+c} \,] = 1 + 0.1 + (0.2 \times 0.1) + (0.8 \times 0.1)$

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:

$$\hat{\theta}^{(1)}_{+a|+c} = \frac{E_\Theta[N_{+a|+c}]}{E_\Theta[N_{+a|+c}] + E_\Theta[N_{-a|+c}]} = \frac{(0.3\times0.1)+(0.3\times0.9)}{[(0.3\times0.1)+(0.3\times0.9)]+[1+(0.1+0.9)+(0.7\times0.1)+(0.7\times0.9)]} = 0.1$$

$$\hat{\theta}^{(1)}_{+c} = \frac{E_\theta[N_{+c}]}{E_\theta[N_{+c}] + E_\theta[N_{-c}]} = \frac{1.0+(1.0)+(1.0)}{4} = 0.75$$

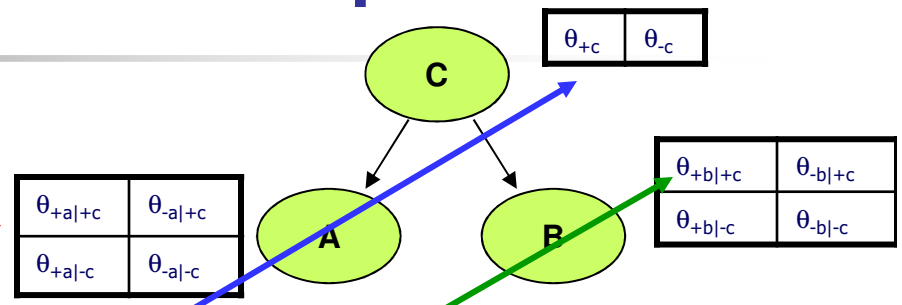
$$\hat{\theta}^{(1)}_{+b|+c} = \frac{E_\theta[N_{+b|+c}]}{E_\theta[N_{+b|+c}] + E_\theta[N_{-b|+c}]} = \frac{0.9+(0.2\times0.9)+(0.8\times0.9)}{3} = 0.6$$

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:

$$\hat{\theta}^{(1)}_{+a|+c} = \frac{E_\Theta[N_{+a|+c}]}{E_\Theta[N_{+a|+c}] + E_\Theta[N_{-a|+c}]} = \frac{(0.3\times0.1)+(0.3\times0.9)}{[(0.3\times0.1)+(0.3\times0.9)]+[1+(0.1+0.9)+(0.7\times0.1)+(0.7\times0.9)]} = 0.1$$

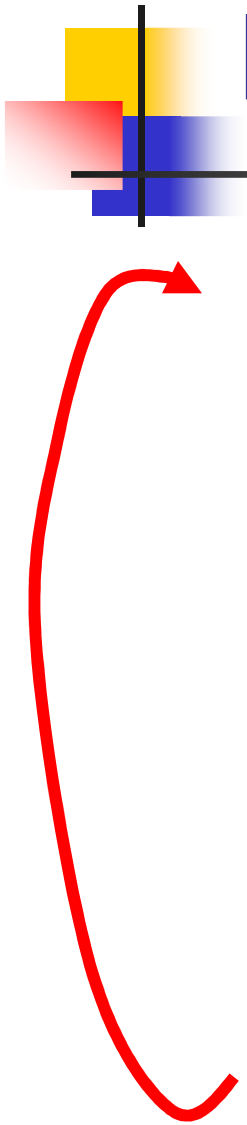
$$\hat{\theta}^{(1)}_{+c} = \frac{E_\theta[N_{+c}]}{E_\theta[N_{+c}] + E_\theta[N_{-c}]} = \frac{1.0+(1.0)+(1.0)}{4} = 0.75$$

$$\hat{\theta}^{(1)}_{+b|+c} = \frac{E_\theta[N_{+b|+c}]}{E_\theta[N_{+b|+c}] + E_\theta[N_{-b|+c}]} =$$

Then

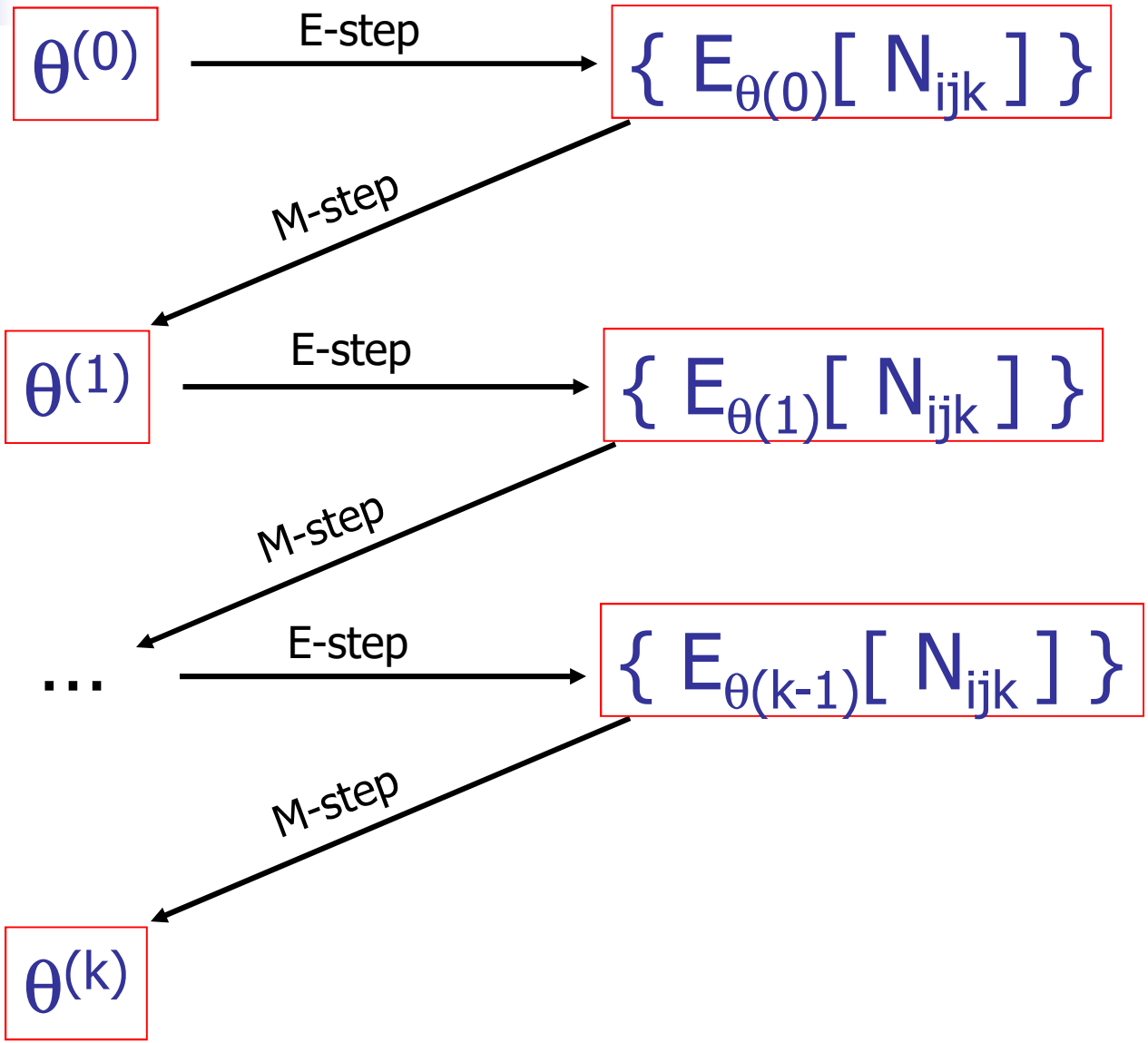
- **E-step**: estimate expected sufficient statistics (wrt missing values) using current $\theta^{(t)}$ values
- **M-step**: compute new $\theta^{(t+1)}$ values, using these expected sufficient statistics

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 Process


θ(0)
E-step
{ Eθ(0)[Nijk] }
M-step
θ(1)
E-step
{ Eθ(1)[Nijk] }
M-step
...
E-step
{ Eθ(k-1)[Nijk] }
M-step
θ(k)

EM Approach

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

- Iteratively, m = 0 ..

  E step: Compute EXPECTED value of N_{ijk}, given $\langle G, \theta^m \rangle$

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

  M step: Update values of θ^{m+1} based on $\hat{N}_{ijk}$

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

  . . . until $|\theta^{m+1} - \theta^m| \approx 0.$

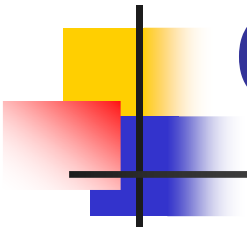
- Return θ^m

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

Facts about EM …

- Converges eventually
- When not converged: Always improves 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 Clique-tree or ClusterGraph
    … just needs one pass
    (as N_{ijk} depends on node+parents)



Outline

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

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 $d_{ij}^{(0)} = \text{Random}[\ \text{Domain}(X_i)\ ]$

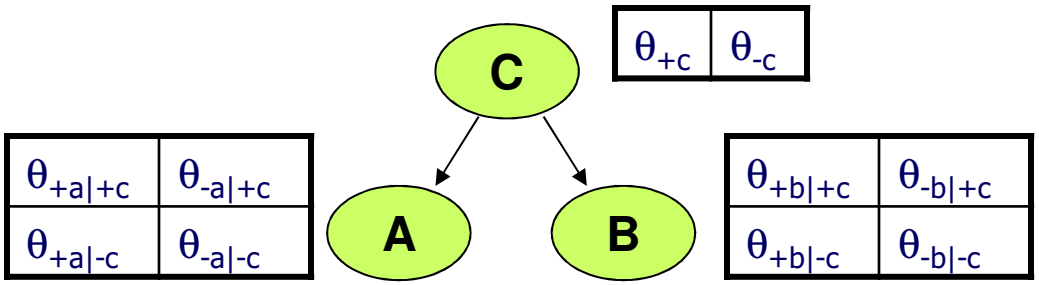
- For $k = 0..$
  - Compute $\Theta^{(k)}$ from $S^{(k)}$ [frequencies]
  - Form $S^{(k+1)}$ by...
    - If $c_{ij} \neq *$, $d_{ij}^{k+1} := c_{ij}$
    - If $c_{ij} = *$ then
      Let d_{ij}^{k+1} be random value for X_i,
      based on current distr Θ^k over $Z - X_i$

- 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



New $S^{(1)}$ =

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

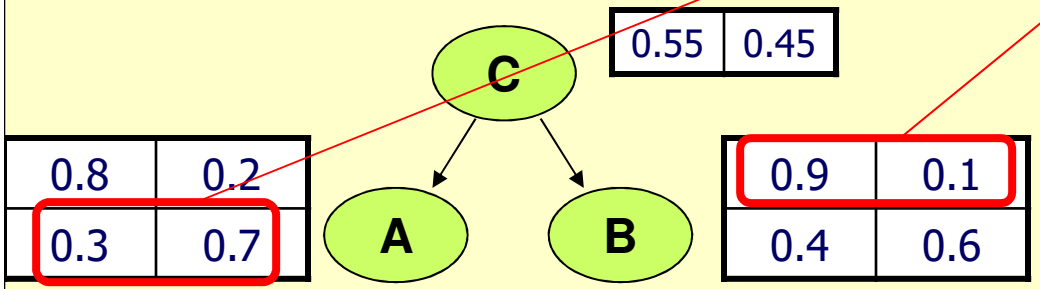
Flip 0.3-coin:

Flip 0.9-coin:

Flip 0.8-coin:

Flip 0.9-coin:

Guess initial values θ^0



Then

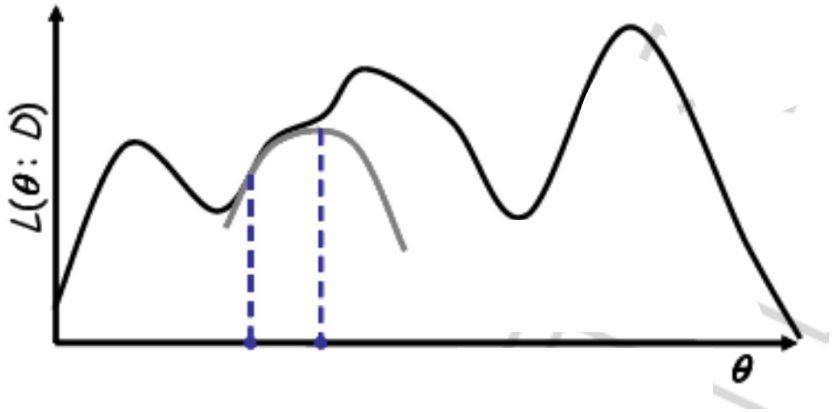
- Use $S^{(1)}$ to get new $\theta^{(2)}$ parameters
- Form new $S^{(2)}$ by drawing new values from $\theta^{(2)}$

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 has converged
- Comparison: Gibbs vs EM
  - + EM "splits" each instance
    …into 2^r parts if r *'s
  - – EM knows when it is done, and what to return

General Issues

- All alg's are heuristic…
  - Starting values $\theta^{(0)}$
  - 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
-
-