HTF: --
B: 8.3 – 8.4
(KF, Chapter 14 – 14.4; RN, Chapter 20)

Learning Belief Net Parameters

R Greiner
Cmput 466 / 551

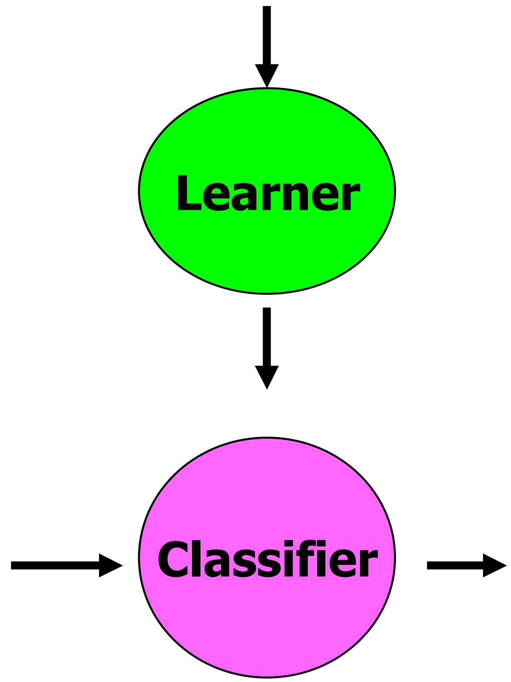
Some material taken from C Guesterin (CMU)

Outline

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

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

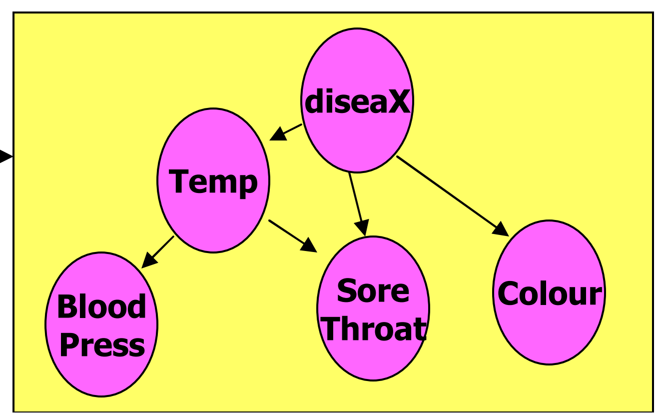
Temp.	Blood Press.	Sore Throat	…	Colour	diseaseX
35	95	Y	…	Pale	No
22	110	N	…	Clear	Yes
:	:			:	:
10	87	N	…	Pale	No

Learner

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

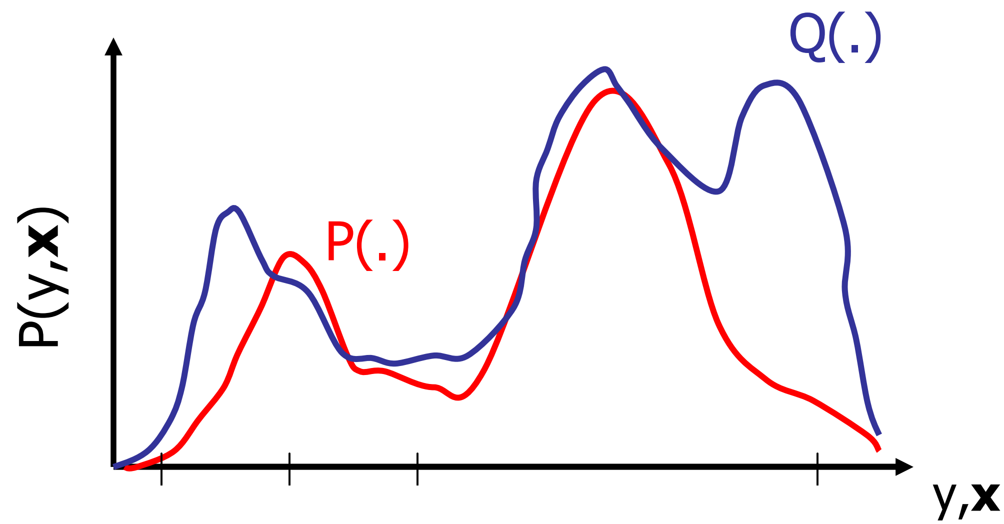
Why Learn Bayes Nets?

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

"Density Estimation"

Generative vs Discriminative

- Generative Learning:
  - Given (sample of) distribution, $P(y,\mathbf{x})$
  - Seek model $Q(y,\mathbf{x})$ that matches $P(y,\mathbf{x})$



- Discriminative Learning:
  - Given (sample of) distribution, $P(y,\mathbf{x})$
  - Seek model $Q(y \mid \mathbf{x})$ that matches $P(y \mid \mathbf{x})$

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

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

- 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 $\mathcal{D}$

$$\bullet\ BN^* = \underset{BN}{\text{argmin}}\ KL(\mathcal{D};\ BN\ )$$

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

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

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

as S drawn from D

$$= \underset{BN}{\text{argmax}} \prod_{d \in D} P_{BN}(\,d\,) = \underset{BN}{\text{argmax}}\ P_{BN}(\,S\,)$$

Best Distribution

- If goal is

  BN that approximates $\mathcal{D}$:

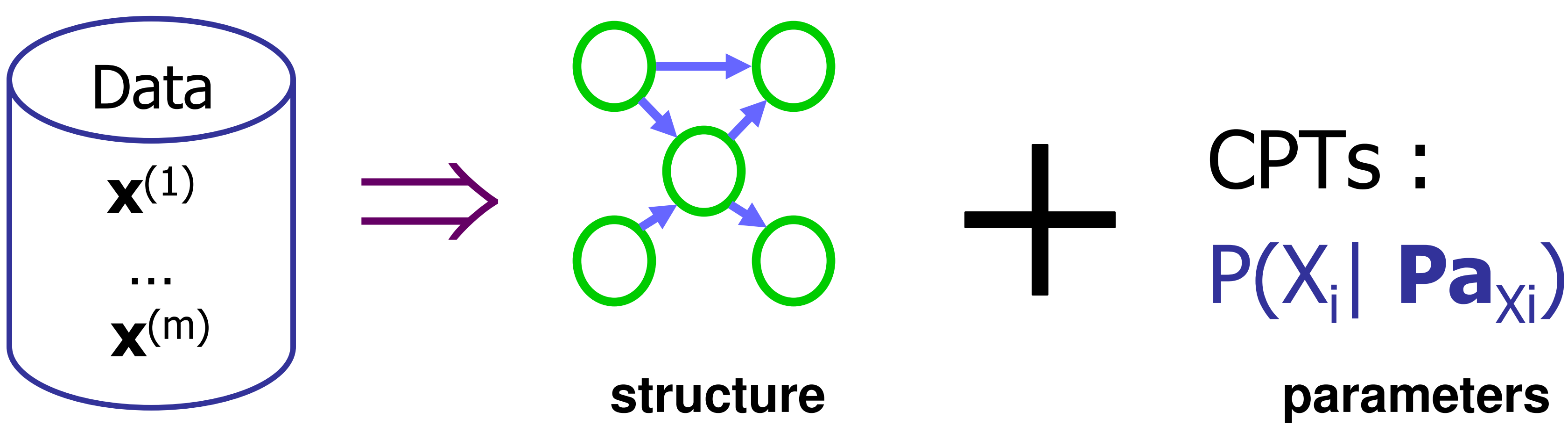
  Find *BN** that maximizes likelihood of data *S*

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

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

Learning Bayes Nets

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 CPtable are independent

   $\theta_A \perp \theta_B$

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

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

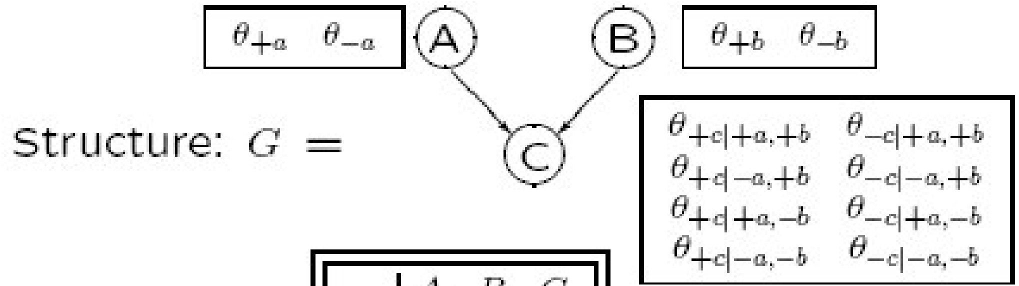
a	$P(B = + \mid A = a)$	$P(B = - \mid A = a)$
$+$	$\theta_{+b\|+a}$	$\theta_{-b\|+a}$
$-$	$\theta_{+b\|-a}$	$\theta_{-b\|-a}$

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

Bayesian Model (items 3–4)

- Later: relax Assumptions 1,2,4

Learning the CPTs



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
  - over discrete variables X_i
  - Complete instances
- $\widehat{\theta}$ = "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????

REPEAT!!

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
- Sequence *S* of α_H Heads and α_T Tails

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

Maximum Likelihood Estimation

- **Data:** Observed set *D* 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:

$$\widehat{\theta} = \arg\max_{\theta} P(\mathcal{D} \mid \theta)$$

$$= \arg\max_{\theta} \ln P(\mathcal{D} \mid \theta)$$

Simple "Learning" Algorithm

$$\widehat{\theta} = \arg\max_{\theta} \ \ln P(\mathcal{D} \mid \theta)$$
$$= \arg\max_{\theta} \ \ln \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

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

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

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

So just average!!!

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

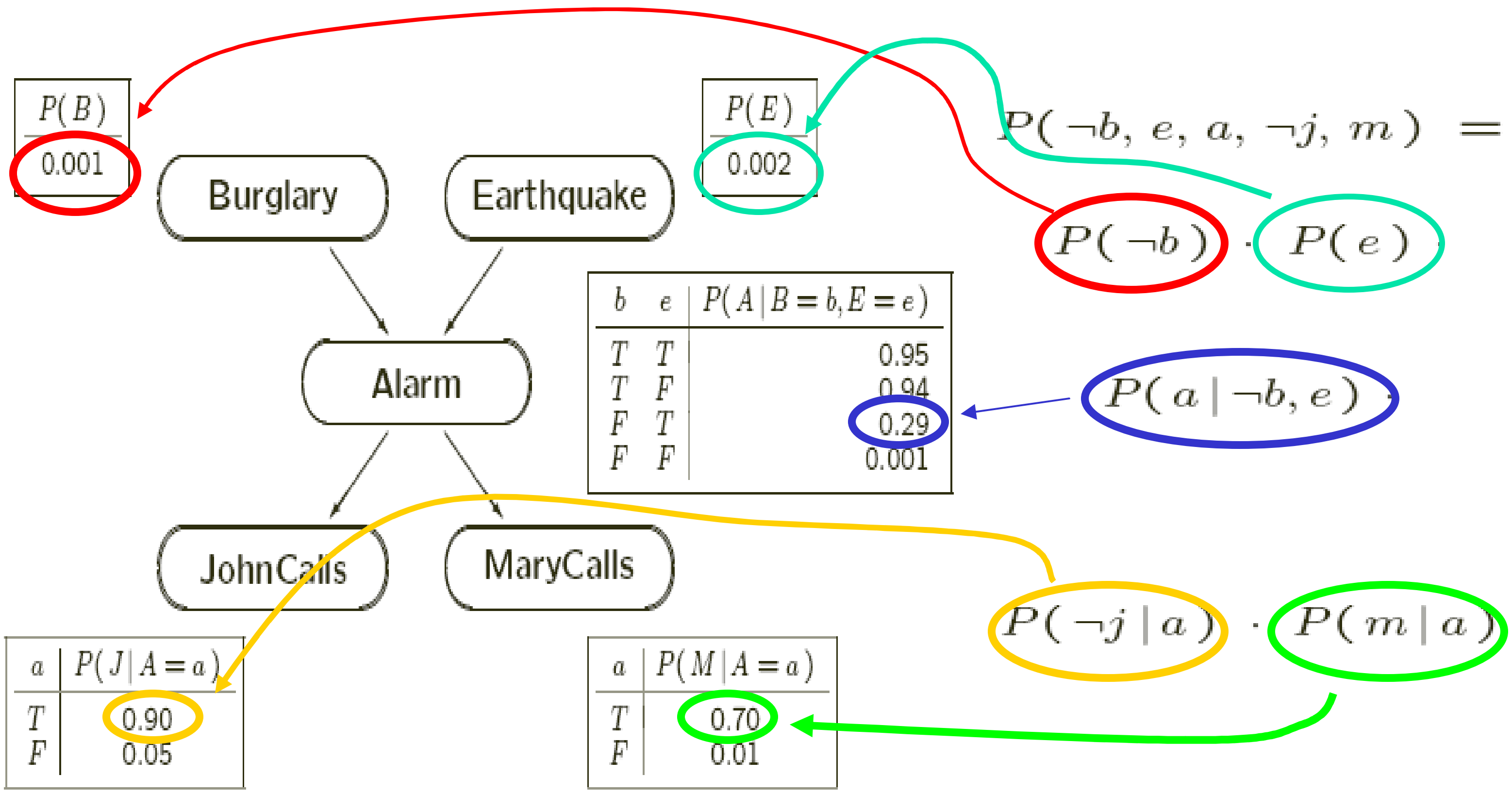
$$P(\neg b, e, a, \neg j, m) =$$

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

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

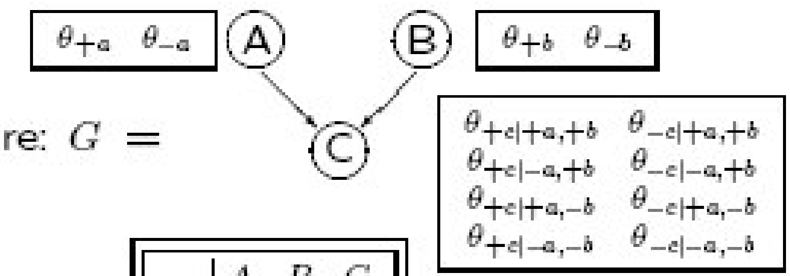
$$0.99 \times 0.02 \times 0.29 \times 0.1 \times 0.70$$

Node independent of predecessors, given parents



Likelihood of the Data (Frequentist)

Given: Structure: G =



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

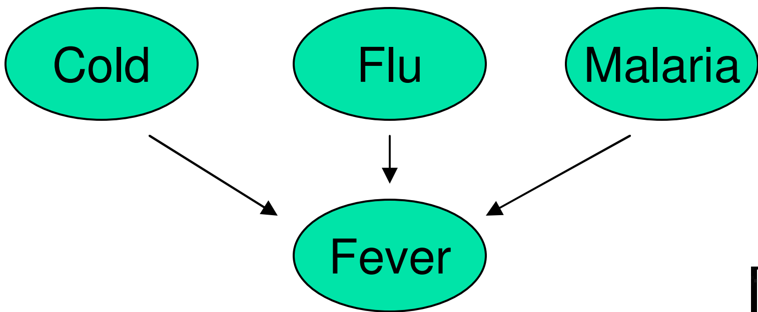
- $P(S|\Theta) = \prod_r P(d_r|\Theta)$

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

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

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

Example of Parameter θ_{ijk}



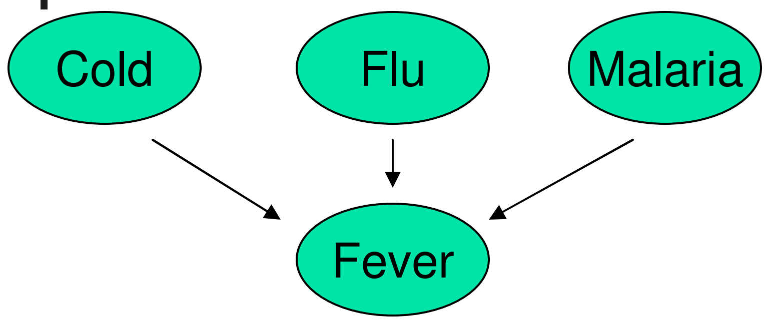
2nd ⇓

4th ⇒

Cold	Flu	Malaria	$P(Fever = ? \mid$ 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}	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 Parameter N_{ijk}



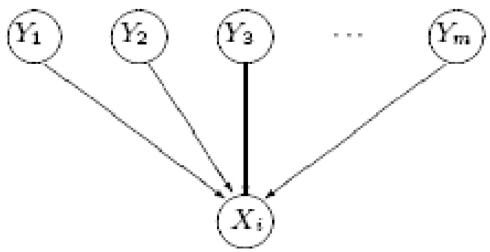
2nd ⇓

4th ⇒

			P(Fever =? | 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}	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	$\ldots$	Y_m	v_{i1}	$\cdots$	v_{ik}	$\cdots$	v_{ir_i}
	u_{11}	u_{21}	$\ldots$	u_{m1}	θ_{111}		θ_{11k}		θ_{11r_i}
	u_{11}	u_{21}	$\ldots$	u_{m2}	θ_{121}		θ_{12k}		θ_{12r_i}
	$\vdots$	$\vdots$	$\ldots$	$\vdots$					
$j^{th} \rightarrow$	$u_{1\ell}$	$u_{2\ell'}$	$\ldots$	$u_{m\ell''}$			θ_{ijk}		
	$\vdots$	$\vdots$	$\ldots$	$\vdots$					
	u_{1r_1}	u_{2r_2}	$\ldots$	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	$\ldots$	Y_m	v_{i1}	$\cdots$	v_{ik}	$\cdots$	v_{ir_i}
	u_{11}	u_{21}	$\ldots$	u_{m1}	N_{111}		N_{11k}		N_{11r_i}
	u_{11}	u_{21}	$\ldots$	u_{m2}	N_{121}		N_{12k}		N_{12r_i}
	$\vdots$	$\vdots$	$\ldots$	$\vdots$					
$j^{th} \rightarrow$	$u_{1\ell}$	$u_{2\ell'}$	$\ldots$	$u_{m\ell''}$			N_{ijk}		
	$\vdots$	$\vdots$	$\ldots$	$\vdots$					
	u_{1r_1}	u_{2r_2}	$\ldots$	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 $S = \{ c_r \}$, ...

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

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

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

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

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

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

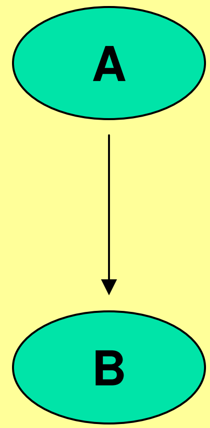
- Observed Frequency Estimates !
- Undefined if $\sum_k N_{ijk} = 0$ … $\#(\mathbf{Pa}_i = \mathbf{pa}_{i,j})=0$

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 $[\,X_i = v_{ik},\ Pa_i = pa_{ij}]$, N_{ijk} += 1
- Return parameters: $$\theta_{ijk} = \frac{N_{ijk}}{\sum_r N_{ijr}}$$

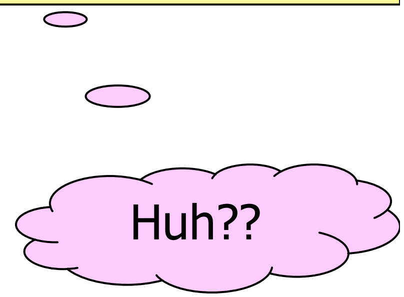
Example



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

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

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



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

≈ 50/50 …

Bayesian Learning

Repeat!

- Use Bayes rule:

$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta) P(\theta)}{P(\mathcal{D})}$$

likelihood → $P(\mathcal{D} \mid \theta)$; *prior* → $P(\theta)$; *posterior* → $P(\theta \mid \mathcal{D})$

- Or equivalently:

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

Repeat!

Bayesian Learning

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

posterior *likelihood* *prior*

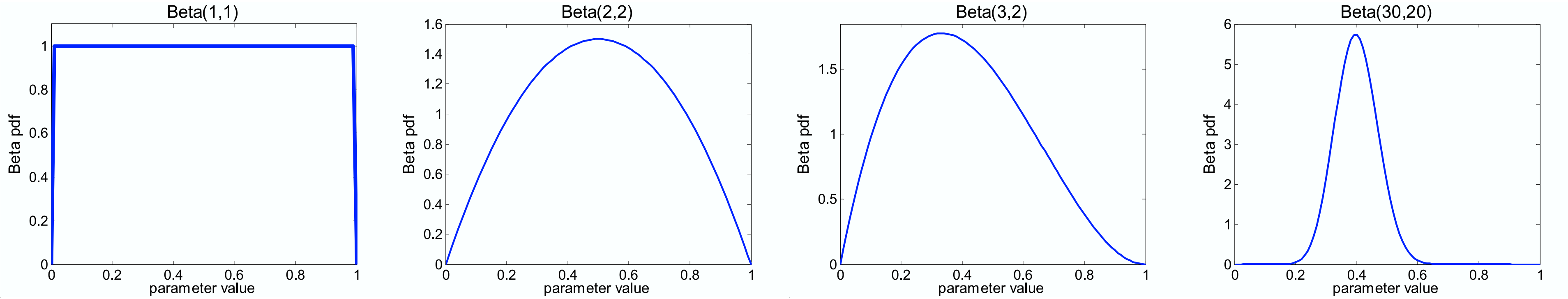
- Likelihood function is simply Binomial:

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

- 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)

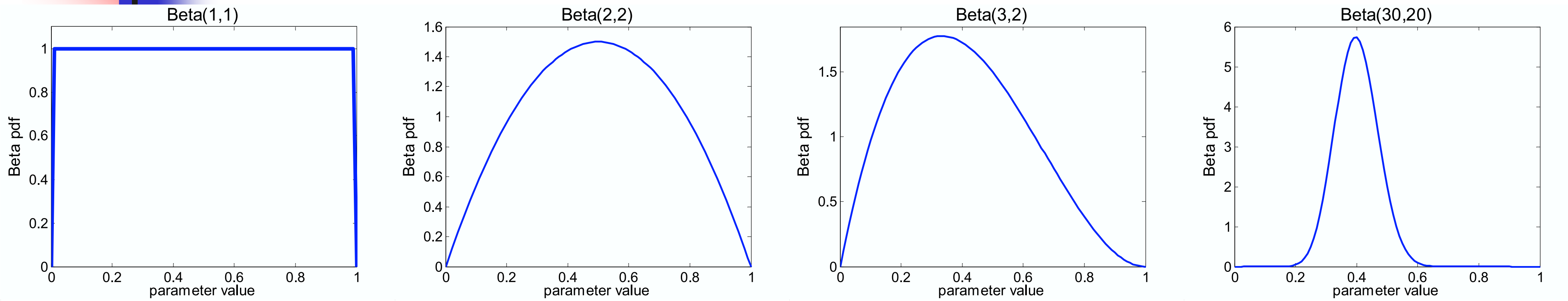
Repeat!



- 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: (a-1) / (a+b-2)
  - Variance: $a \times b / [(a+b)^2 (a+b-1)]$

Posterior distribution… from Beta

Repeat!



Prior $P(\theta)$

Likelihood $P(D|\theta)$

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

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

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

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

Same form! Conjugate!

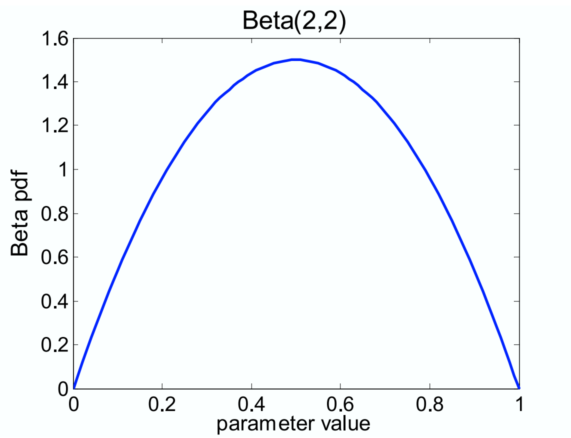
Repeat!

Posterior Distribution

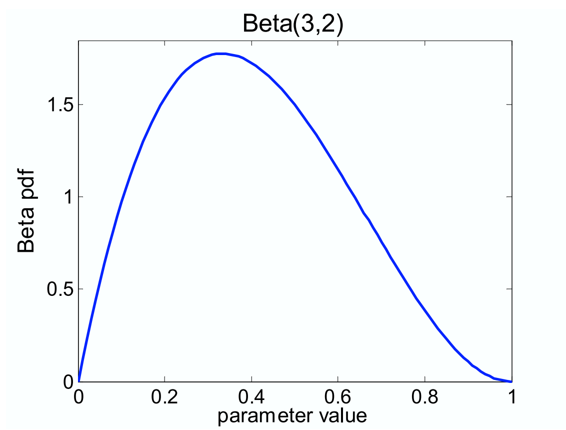
- Prior: $\theta \sim \mathrm{Beta}(\alpha_H, \alpha_T)$
- Data $\mathcal{S}$: m_H heads, m_T tails

- Posterior distribution:

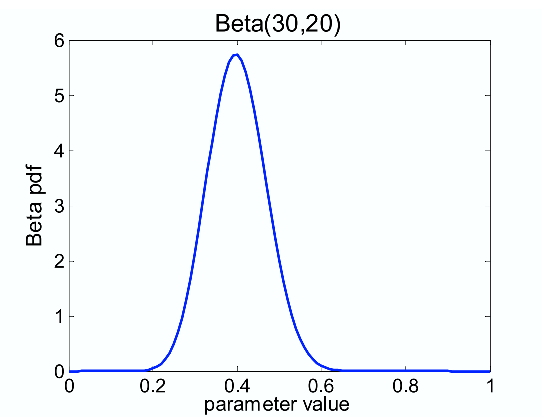
  $\theta \mid \mathcal{S} \sim \mathrm{Beta}(\, m_H + \alpha_H,\ m_T + \alpha_T \,)$



Prior

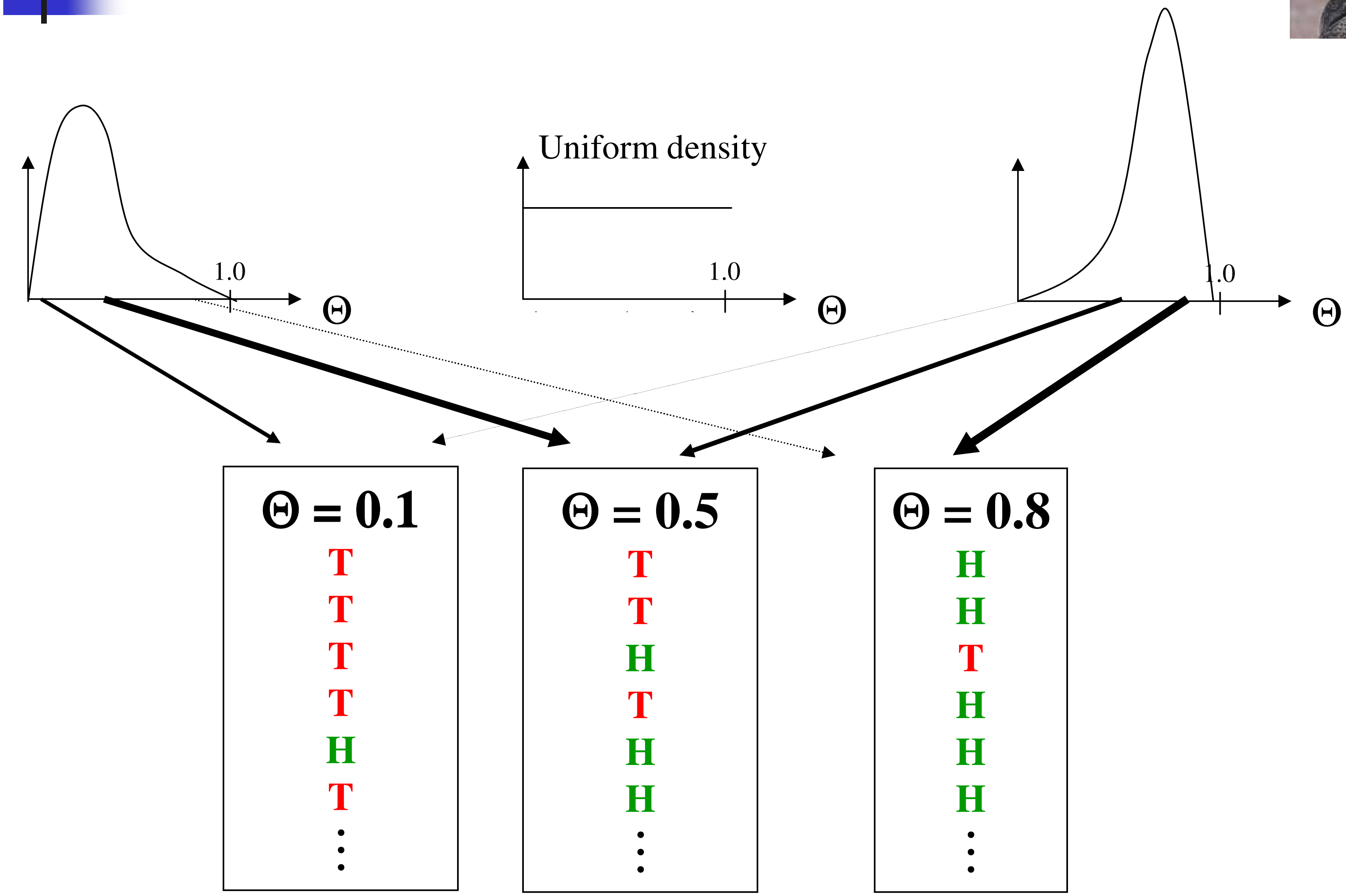


+ observe 1 head

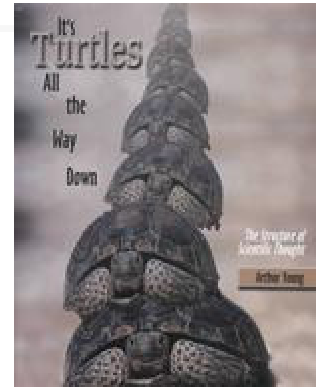


+ observe 27 more heads; 18 tails

Two (related) Distributions: Parameter, Instances



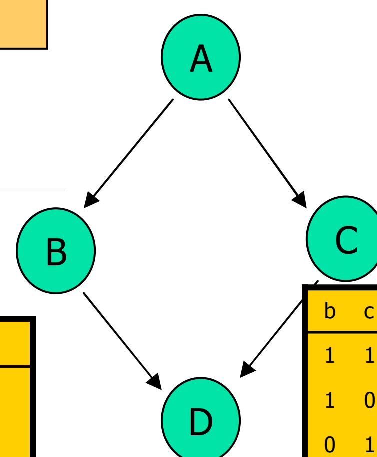
Distribution over Parameter



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

only a distribution!

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

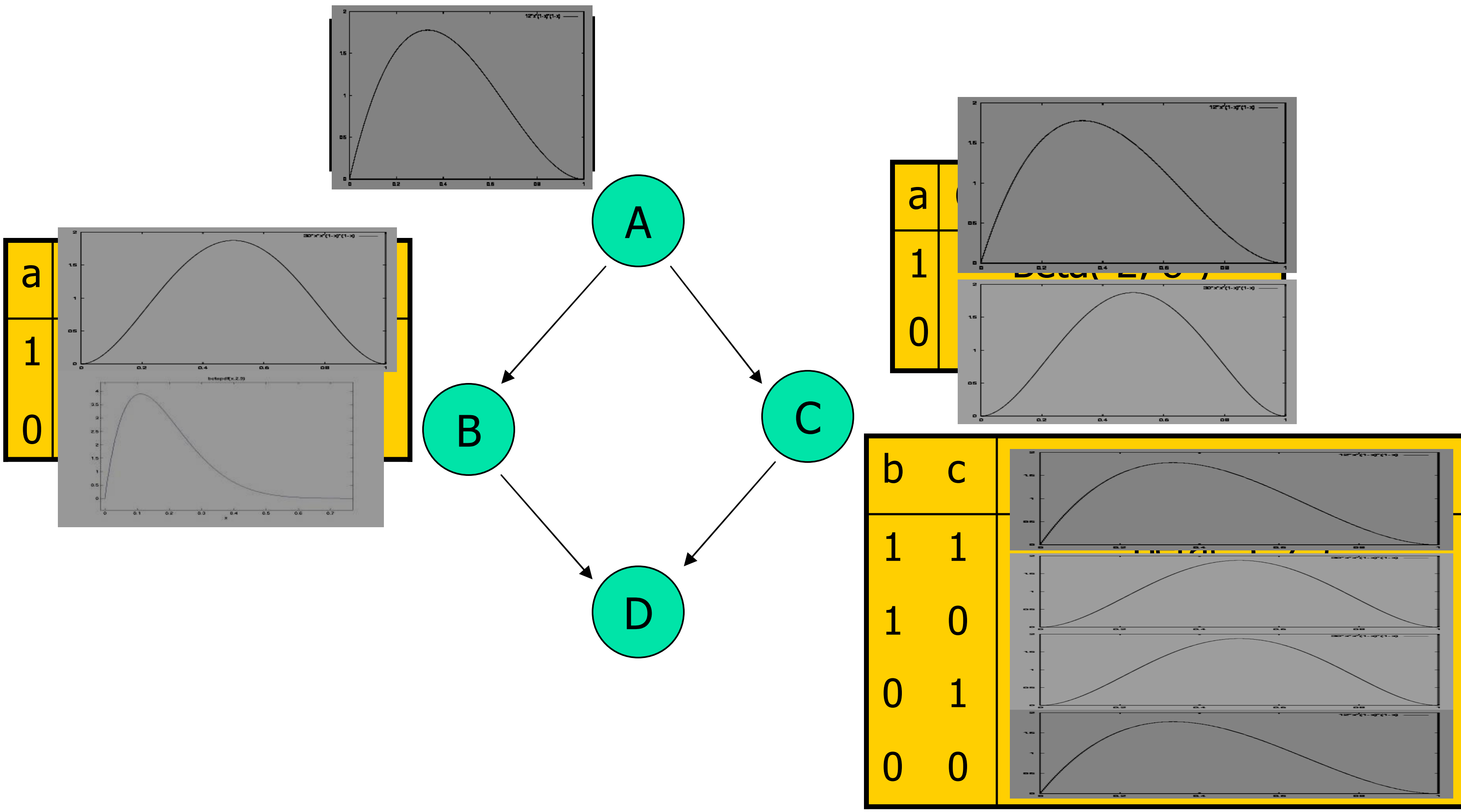


a	$\theta_{C=1\|A=a}$	$\theta_{C=0\|A=a}$
1	0.200	0.800
0	0.367	0.633

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*

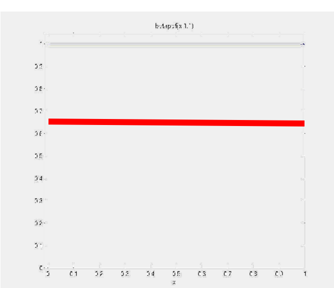
- $\theta_{B|A=1} = \langle \theta_{B=0|A=1},\ \theta_{B=1|A=1} \rangle \sim$ Beta(1, 1)

  kinda like seeing 2 instances with ⟨*A=1* ⟩:

1 with ⟨*A=1, B=0*⟩ →

1 with ⟨*A=1, B=1*⟩ →

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)

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

- 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)

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

- Then observe data *S*
  - *2* $\langle A=1, B=1 \rangle$
  - *4* $\langle A=1, B=0 \rangle$

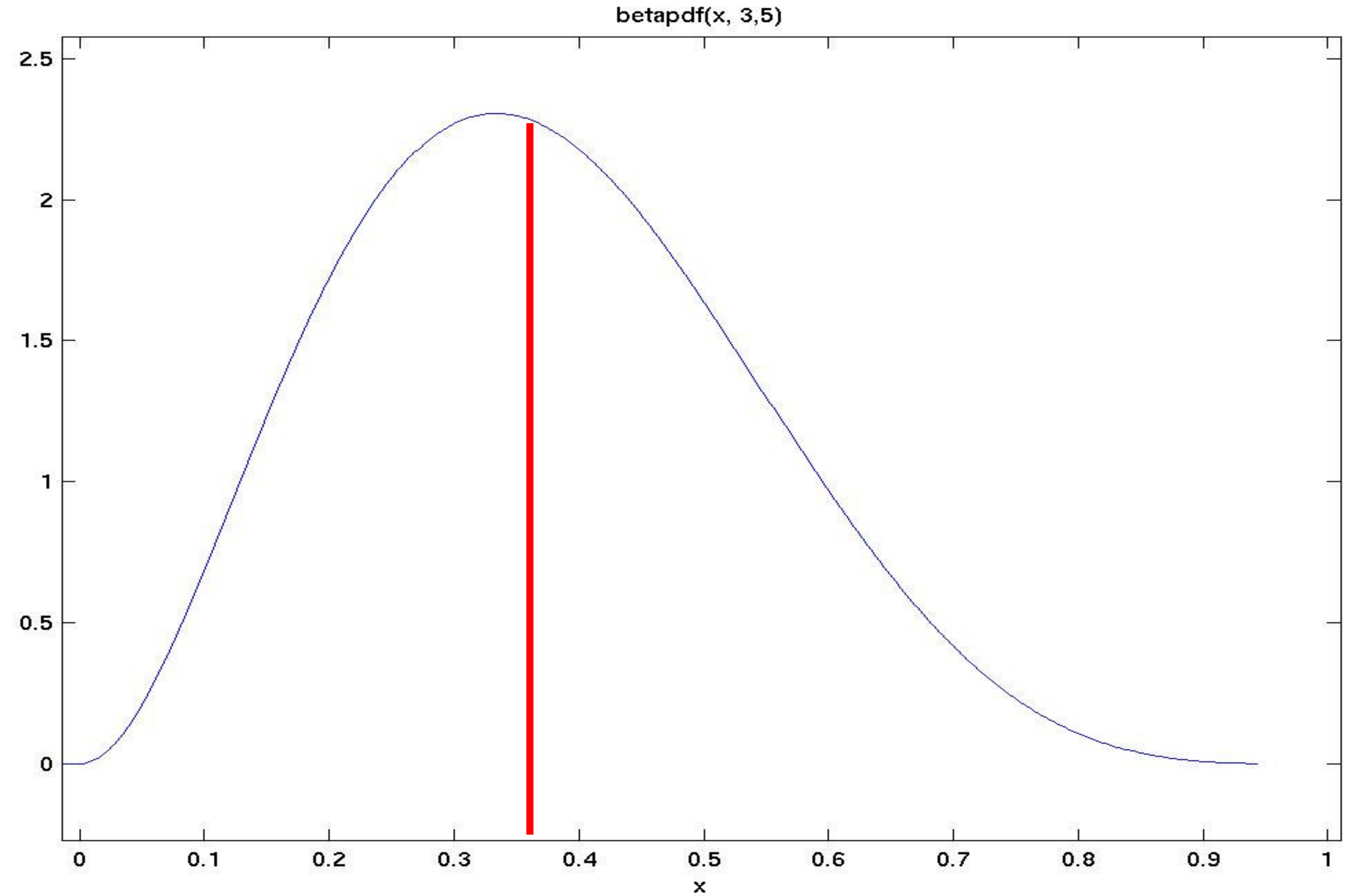
A	B	C	E
1	1	0	1
1	1	1	1
1	0	1	0
1	0	1	0
1	0	0	0
1	0	0	1
0	0	0	1
⋮	⋮	⋮	⋮

- *New distribution is*

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

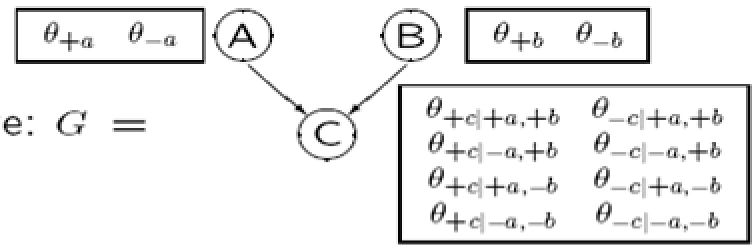
$$\Rightarrow \quad E[\theta_{B=1|A=1} \mid S] = \widehat{\theta}_{+b|+a} \mid S = \frac{3}{3+5} = 0.375$$

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

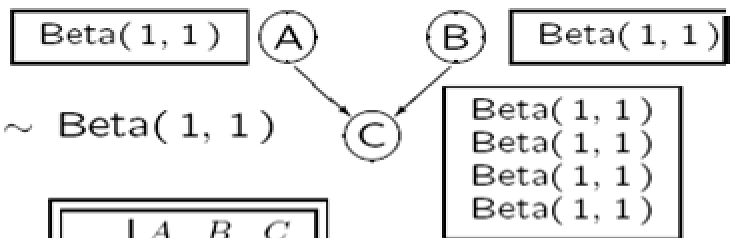


Posterior Distribution of Θ

Given: Structure: G =



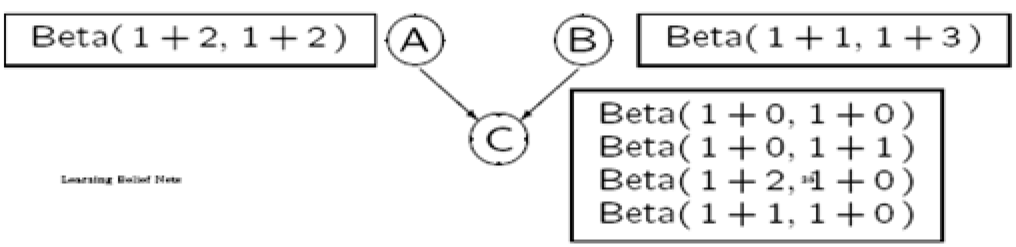
where $\Theta_{X|y} \sim$ Beta(1, 1)



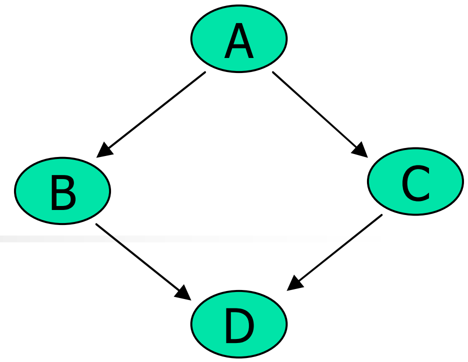
- Given sample S =

	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. . .



Posterior Distribution



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

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

- Compare to Frequentist:

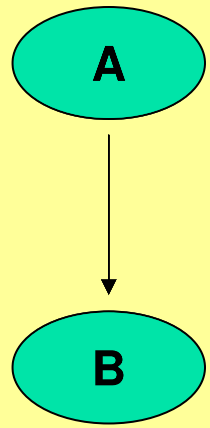
$$\hat{\theta}_{ijk} = \frac{N_{ijk}}{\sum_r N_{ijr}}$$

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 $[\,X_i = v_{ik},\ Pa_i = pa_{ij}]$, u_{ijk} += 1
- Set parameters:
  $\theta_{ij} \mid S \sim Dir(u_{ij1}, \ldots, u_{ijr})$
- If want expected value: $$E[\theta_{ijk}] = \frac{u_{ijk}}{\sum_r u_{ijr}}$$

Example



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

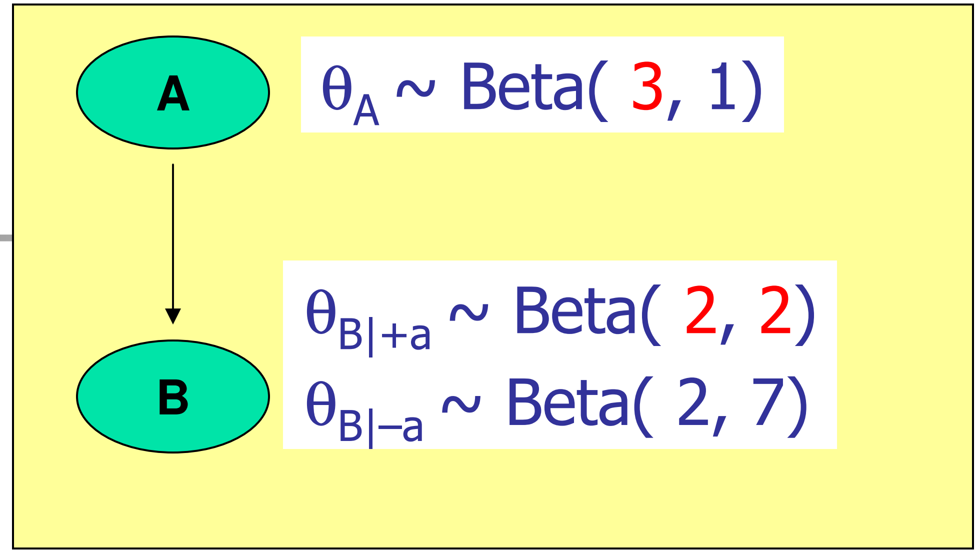
$\theta_{B|+a} \sim \text{Beta}(\alpha_{+b|+a}, \alpha_{-b|+a})$

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

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

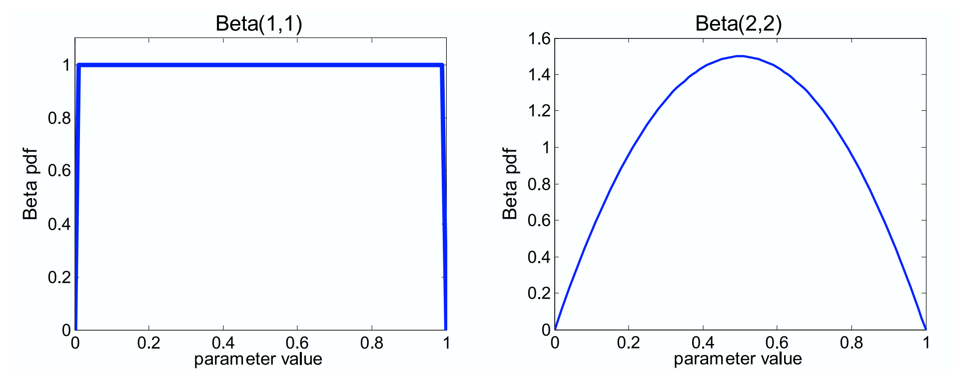
A	B
+	+
+	−

Example



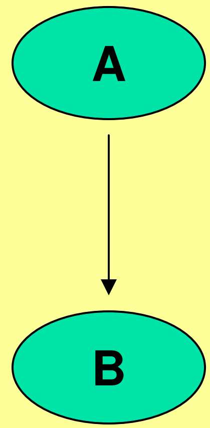
- Buckets
  - N_{+a} := ~~1~~ ~~2~~ 3
  - N_{-a} := 1
  - $N_{+b|+a}$:= ~~1~~ 2
  - $N_{-b|+a}$:= ~~1~~ 2
  - $N_{+b|-a}$:= 2
  - $N_{-b|-a}$:= 7

A	B
+	+
+	−



Example

If you want POINT estimates…



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

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

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

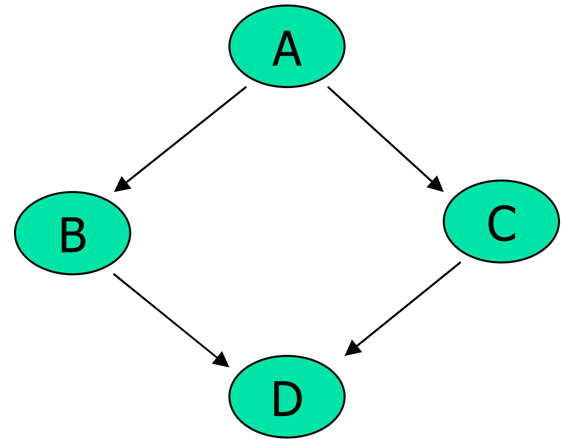
- Buckets
  - N_{+a} := ~~1~~ ~~2~~ 3
  - N_{-a} := 1
  - $N_{+b|+a}$:= ~~1~~ 2
  - $N_{-b|+a}$:= ~~1~~ 2
  - $N_{+b|-a}$:= 2
  - $N_{-b|-a}$:= 7

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

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

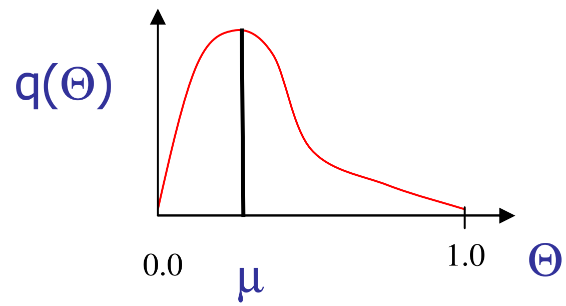
  is function of parameters Θ
- Eg…

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

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

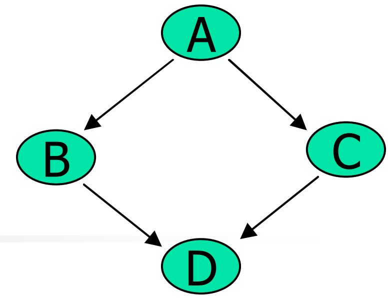
- $P_\Theta(C=c \mid \mathbf{E=e})$ depends on Θ
- As Θ is r.v., so is response

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



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

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



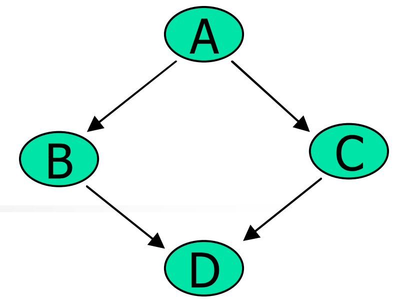
$$q(\Theta) = P_\Theta(A=1 \mid B=1, C=1) = \frac{\theta_{A=1}\,\theta_{B=1|A=1}\,\theta_{C=1|A=1}}{\sum_a \theta_{A=a}\,\theta_{B=1|A=a}\,\theta_{C=1|A=a}}$$

- 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 $q^{(R)} = 1/R \sum_i q(\Theta^{(i)})$

But … easier approach:

- As $R \to \infty$, $q^{(R)} \to E[q]$

Predictive Distribution



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

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

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

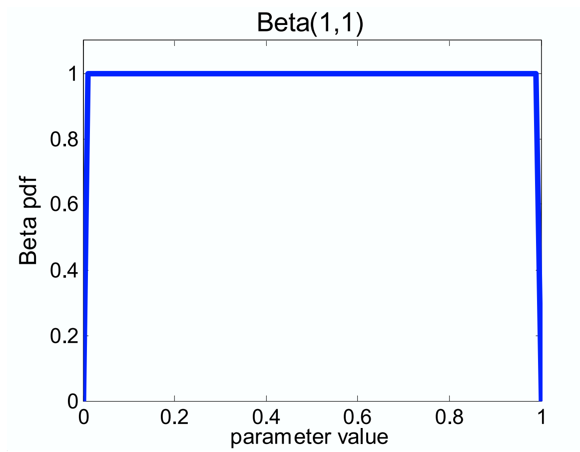
- $BN^{\mathcal{D}} = [\mathcal{G}, \Theta^{\mathcal{D}}]$ with $\quad \theta^{\mathcal{D}} = \left\{ \frac{N_{ijk}+1}{\sum_k (N_{ijk}+1)} \right\}$

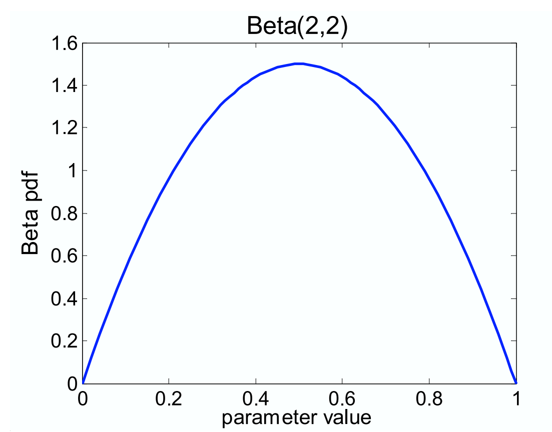
  Compute $E[\ q(\theta)\ ]$ by using just $BN^{\mathcal{D}}$!

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

Alternative "Encoding"

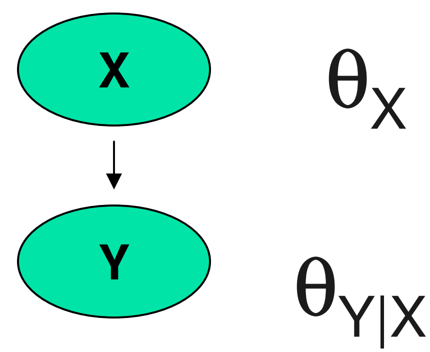
- Beta(a, b) ≡ B(m; μ, 1-μ) where
  - m = (a+b) … effective sample size
  - μ = a/(a+b)
- Eg…
  - Beta(1,1) = B(2; 0.5, 0.5)
  - Beta(10,10) = B(20; 0.5, 0.5)
  - Beta(7, 3) = B(10; 0.7, 0.3)
  - …





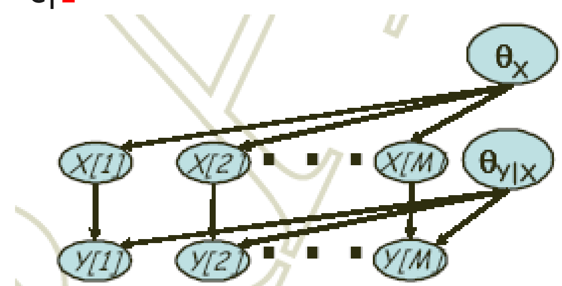
Bayesian Learning for 2-node BN

- Parameters θ_X, $\theta_{Y|X}$
- Priors:
  - $\theta_X \sim \text{Dirichlet}(\alpha_{X=1}, \ldots, \alpha_{X=k})$
  - $P(\theta_{Y|X})$: k different distributions:
    for each x,
    $\theta_{Y|X=x} \sim \text{Dirichlet}(\alpha_{Y=1|x}, \ldots, \alpha_{Y=k|x})$



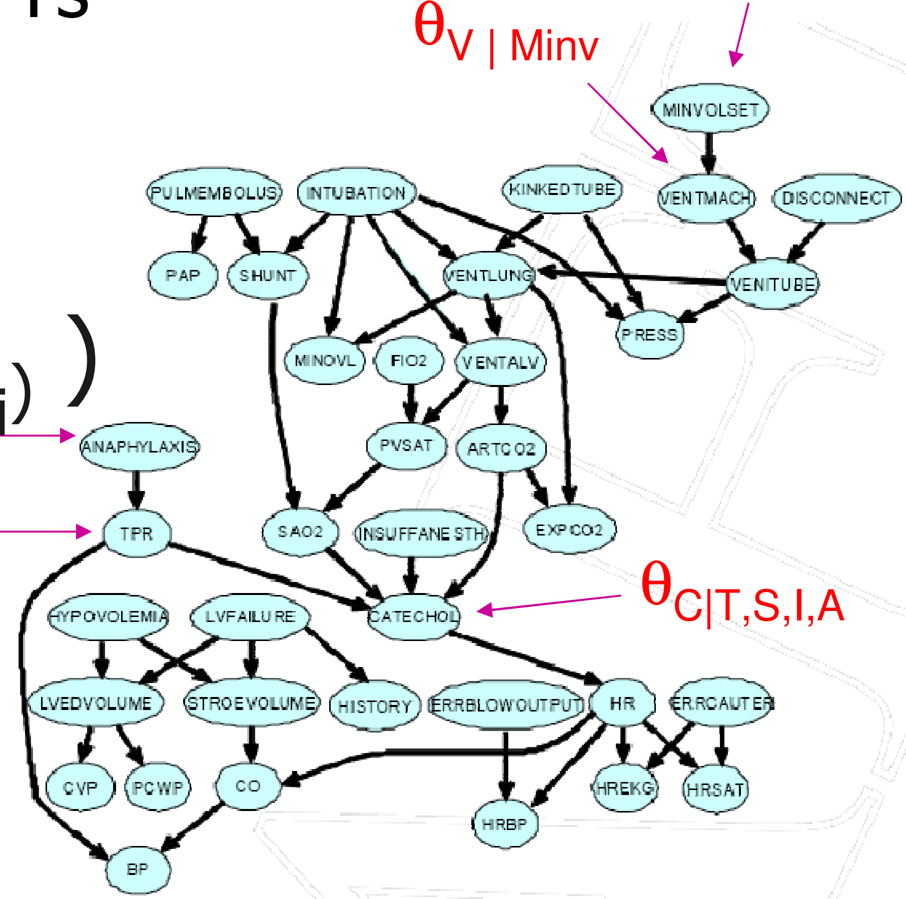
$$\theta_{Y|X=0} \sim \text{Dirichlet}(\alpha_{Y=a|0}, \alpha_{Y=b|0}, \alpha_{Y=c|0})$$
$$\theta_{Y|X=1} \sim \text{Dirichlet}(\alpha_{Y=a|1}, \alpha_{Y=b|1}, \alpha_{Y=c|1})$$

  - Independent
    $\theta_{Y|X=0} \perp \theta_{Y|X=1}$



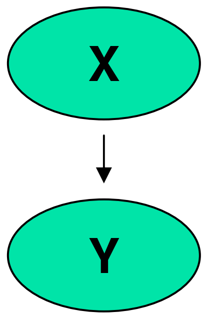
Important Assumption wrt Prior

- **Global parameter independence:**
  - Prior over parameters == product of priors over CPTs
  - $\theta_{ijk} \perp \theta_{rst}$
    - $\theta_{V \mid Minv} \perp \theta_{Minv}$
- $P(\Theta) = \prod_i P(\theta_{X_i \mid Pa(X_i)})$



Global parameter independence, d-separation and local prediction

- Independencies in **meta BN**:



θ_X

$\theta_{Y|X}$

$$\theta_{Y|X} \perp \theta_X$$

- **Proposition**:
  If prior satisfies global parameter independence, then given fully observable data *D*,

$$\theta_{Y|X} \perp \theta_X \mid D$$

$$P(\theta \mid \mathcal{D}) = \prod_i P(\theta_{X_i|\mathbf{Pa}_{X_i}} \mid \mathcal{D})$$

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
    - Gradient Descent
    - EM
    - Gibbs
  - Learning Structure

#2: Known structure, Missing data

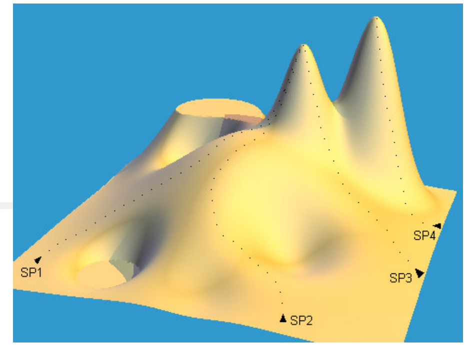
- To find good Θ, need to compute $P(\Theta, \mathcal{D} \mid \mathcal{G})$
- Easy if ..

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

~~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
  - … 1^{st}-order Taylor series e

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

Need derivative!

```
Procedure Gradient-Ascent (
  θ^1,     // Initial starting point
  f_obj,   // Function to be optimized
  δ        // Convergence threshold
)
  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 into constraint region, $[0,1]^{|\Theta|}$
return(Θ)

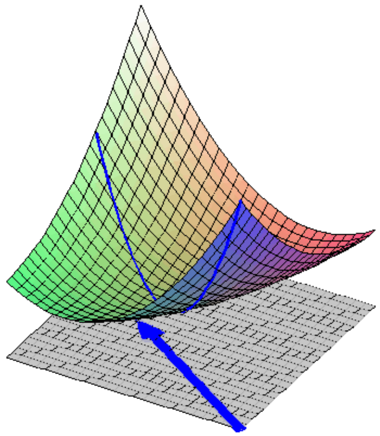
Note: Computed $P(v_{ik}, pa_{ij} \mid c_r)$ to deal with c_r
$\Rightarrow$ can "piggyback" computation

Issues with Gradient Ascent

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

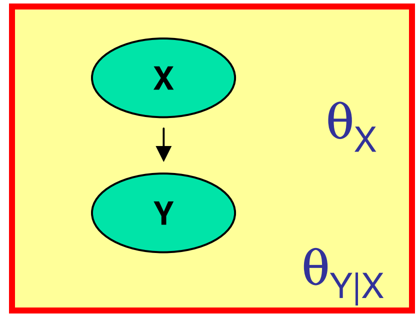
- Lots of Tricks for efficient ascent
  - Line Search
  - Conjugate Gradient
  - …

  [See earlier notes on optimization]



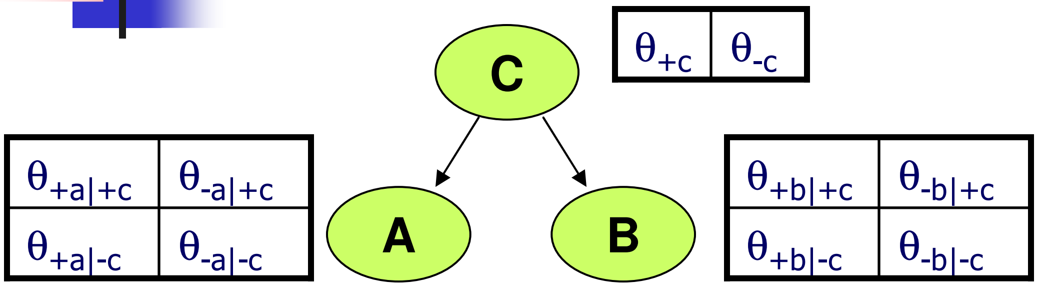
Expectation Maximization (EM)

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

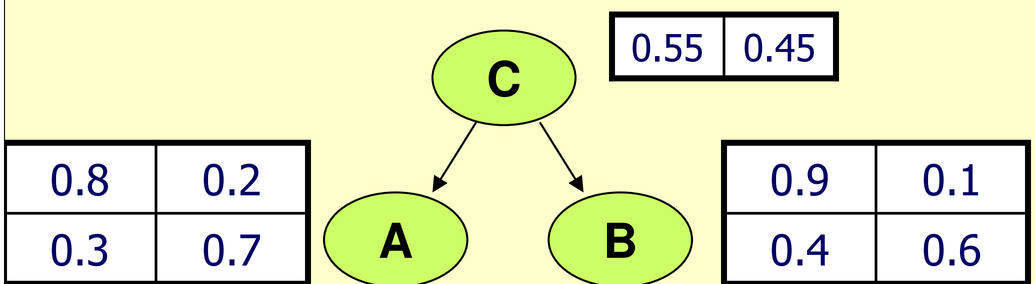
EM Approach – E Step



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

$$E_{\theta}0 [\ N_{+b|+c}\ ] = 0.9+(0.2\times0.9)+(0.8\times0.9)$$

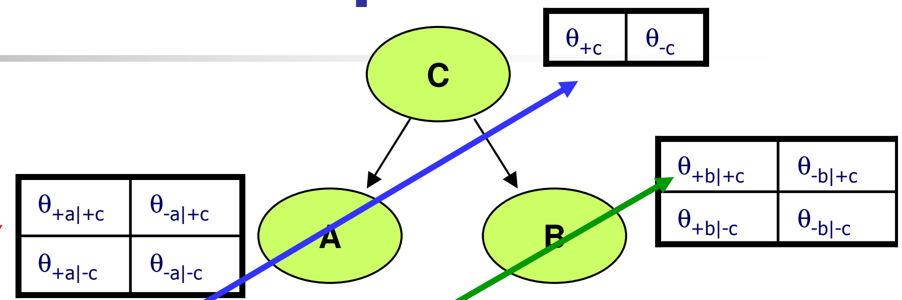
$$E_{\theta}0 [\ N_{-b|+c}\ ] = 0.1+(0.2\times0.1)+(0.8\times0.1)$$

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



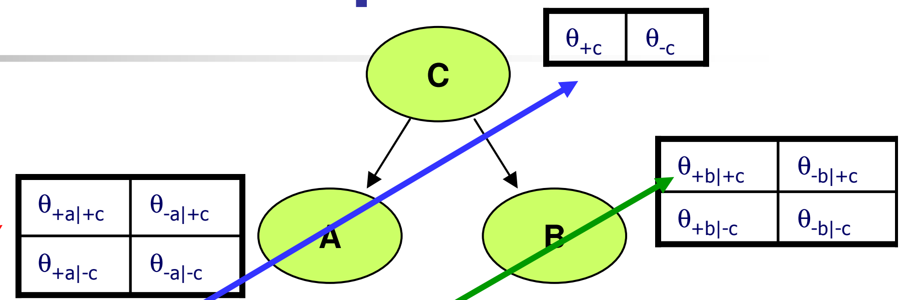
- New estimates:

$$\hat{\theta}^{(1)}_{+a|+c} = \frac{E_{\Theta}[N_{+a|+c}]}{E_{\Theta}[N_{+a|+c}] + E_{\Theta}[N_{-a|+c}]} = \frac{(0.8\times 0.1)+(0.8\times 0.9)}{[(0.8\times 0.1)+(0.8\times 0.9)]+[1+(0.1+0.9)+(0.2\times 0.1)+(0.2\times 0.9)]} = 0.233$$

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

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

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:

$$\hat{\theta}^{(1)}_{+a|+c} = \frac{E_\Theta[N_{+a|+c}]}{E_\Theta[N_{+a|+c}] + E_\Theta[N_{-a|+c}]} = \frac{(0.8\times0.1)+(0.8\times0.9)}{[(0.8\times0.1)+(0.8\times0.9)]+[1+(0.1+0.9)+(0.2\times0.1)+(0.2\times0.9)]} = 0.233$$

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

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

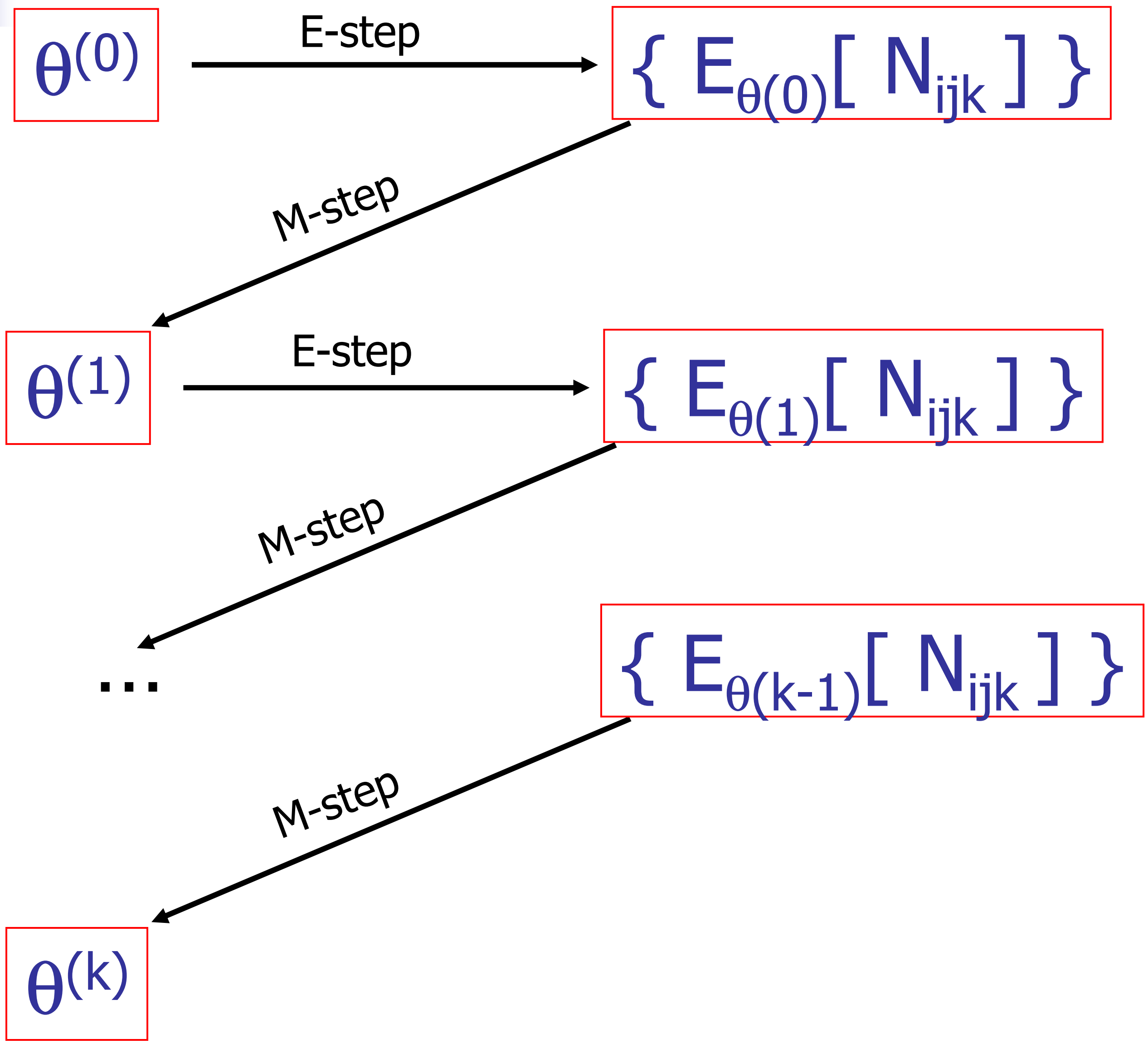
Then

- **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


$\theta^{(0)}$
E-step
$\{ E_{\theta(0)}[N_{ijk}] \}$
M-step
$\theta^{(1)}$
E-step
$\{ E_{\theta(1)}[N_{ijk}] \}$
M-step
…
$\{ E_{\theta(k-1)}[N_{ijk}] \}$
M-step
$\theta^{(k)}$

EM Approach

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

- Iteratively, $k = 0, \ldots$

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

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

  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" $\rightarrow \alpha_{ijk}$
2. Finds local optimum
3. Used for HMM
4. Views each tuple with k "$*$"s as $O(2^k)$ partial-tuples

Facts about EM …

- Converges eventually
- 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)

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

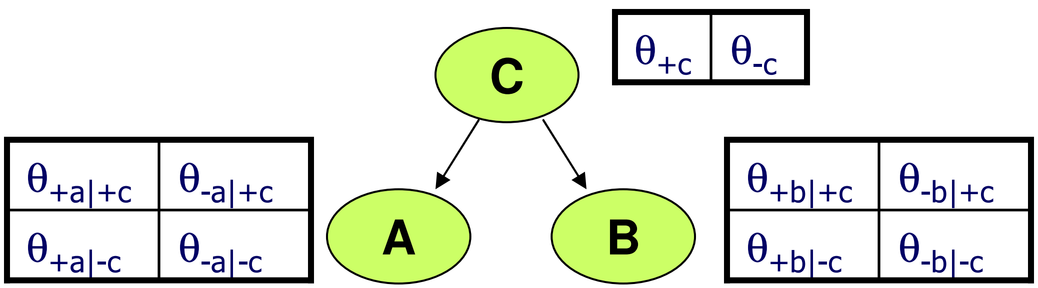
- For $k = 0..$
  - Compute $\Theta^{(k)}$ from $S^{(k)}$ [frequencies]
  - Form $S^{(k+1)}$ by...
    - $d_{ij}^{k+1} = c_{ij}$
    - If $c_{ij} = *$ then
      Let d_{ij}^{k+1} be random value for X_i,
      based on current distr Θ^k over $Z - X_i$

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

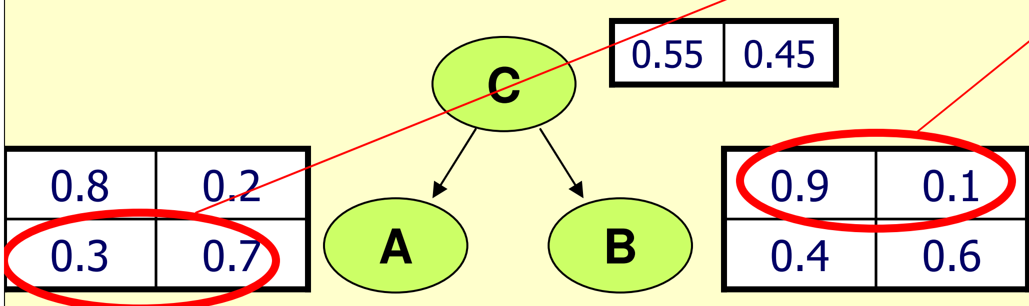
Note: As $\Theta^{(k)}$ based on COMPLETE DATA $S^{(k)}$
$\Rightarrow \Theta^{(k)}$ can be computed efficiently!

"Multiple Imputation"

Gibbs Sampling – Example



Guess initial values θ^0



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:

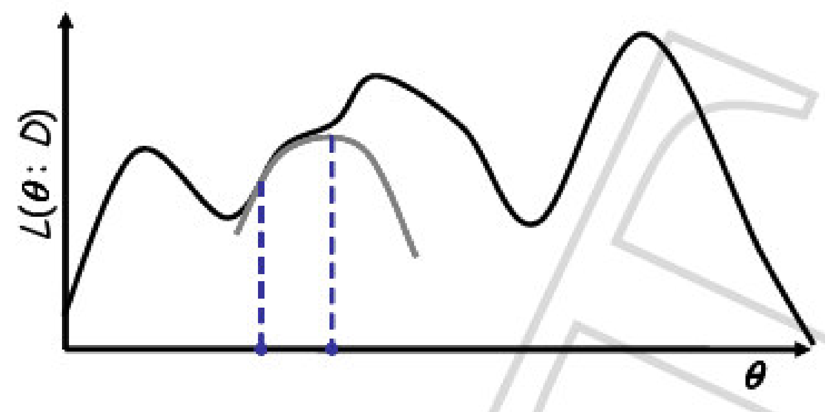
Then
- Use $S^{(1)}$ to get new $\theta^{(2)}$ parameters
- Form new $S^{(2)}$ by drawing new values from $\theta^{(2)}$

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 $\theta^{(0)}$
  - Stopping criteria
  - Escaping local maxima



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

Gaussian Approximation

(Assumes large amounts of data)

- Let $g(\Theta) = \log[P(\,S\,|\,\Theta, \mathrm{G}\,)\;P(\,\Theta\,|\,\mathrm{G}\,)]$
  Let $\tilde{\Theta}_{BN} = \arg\max_{\Theta} g(\Theta)$
  ... also maximizes $P(\,\Theta\,|\,\mathrm{G},\;S\,)$.

  With many samples,
  $\tilde{\Theta}_{BN} \;\approx\; \arg\max_{\Theta}\{P(\,S\,|\,\Theta, \mathrm{G}\,)\}$

- $g(\Theta) \approx g(\tilde{\Theta}_{BN}) - \frac{1}{2}(\Theta - \tilde{\Theta}_{BN})A(\Theta - \tilde{\Theta}_{BN})^t$
  (2^{nd}-order Taylor; A is neg. Hessian of $g(\tilde{\Theta}_{BN})$)

  So...
  $P(\,\Theta\,|\,\mathrm{G},\;S\,) \propto P(\,S\,|\,\Theta, \mathrm{G}\,)\;P(\,\Theta\,|\,\mathrm{G}\,)$
  $\approx$
  $P(\,S\,|\,\tilde{\Theta}_{BN}, \mathrm{G}\,)\;P(\,\tilde{\Theta}_{BN}\,|\,\mathrm{G}\,)e^{\{(\Theta-\tilde{\Theta}_{BN})A(\Theta-\tilde{\Theta}_{BN})^t\}}$

  ... which looks (approximately) Gaussian!

- Now use
  gradient descent or EM

Note: Can often use values computed during inference!

Summary of Approaches

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