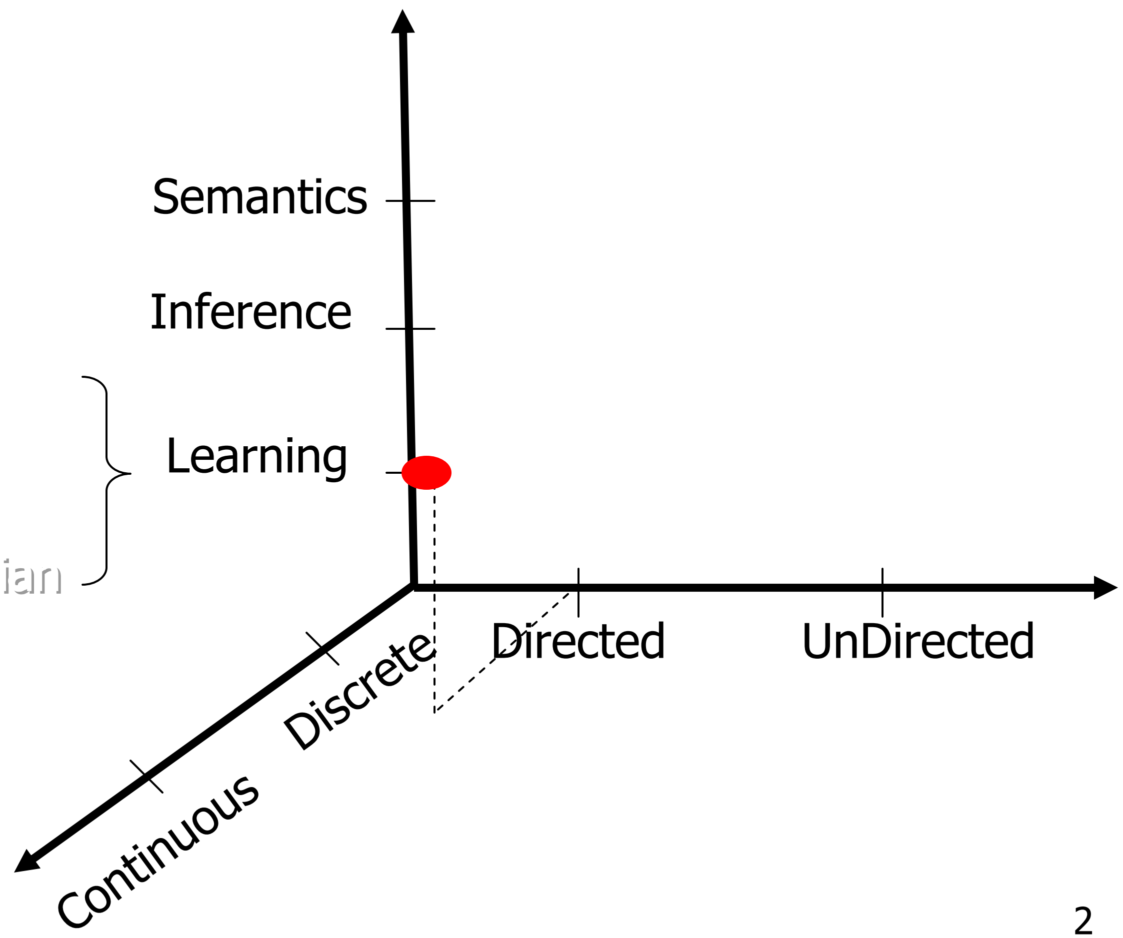
Learning Bayesian Nets Parameters from Partial Data

KF, Chapter 18-18.2

Some material taken from C Guesterin (CMU)

Space of Topics

Learning…
- Parameter, Structure
- Data: Complete, Missing
- Framework: Frequentist, Bayesian



Learning Belief Net Parameters from Partial Data

- Framework
  - Why is the data missing? … MCAR, MAR, …
  - Why more challenging?
- Approaches
  - Gradient Ascent
  - EM
  - Gibbs

Learning from Missing data

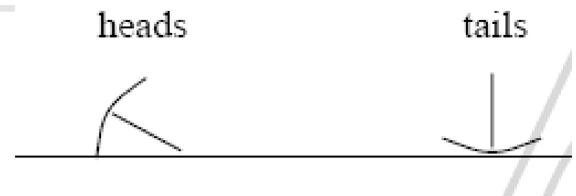
- To find good Θ, need to compute $P(\Theta, S \mid \mathcal{G})$
- Easy if ..

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

~~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, ...

Why is the data missing?



- Estimating P(Heads) = θ
  - Earlier: data = [H, T, H, H, …, T]
  - Now: data = [H, T, ?, ?, H, …, T]
- Thumbtack falls off table, …not recorded
  - No information in "?"

vs

- Recorder doesn't like "Tails", and so omits those values
  - Here, "?" means "Tails" – lots of info!

Formal Model

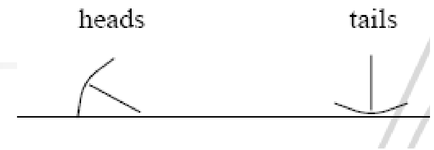
- $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$: set of r.v.s

  $\mathbf{O}_\mathbf{X} = \{O_1, O_2, \ldots, O_n\}$:
  corresponding set of *observability variables*
- $P_{miss}(\mathbf{X}, \mathbf{O}_\mathbf{X}) = P(\mathbf{X}) \cdot P_{miss}(\mathbf{O}_\mathbf{X} \mid \mathbf{X})$
- $P(\mathbf{X})$ parameterized by θ

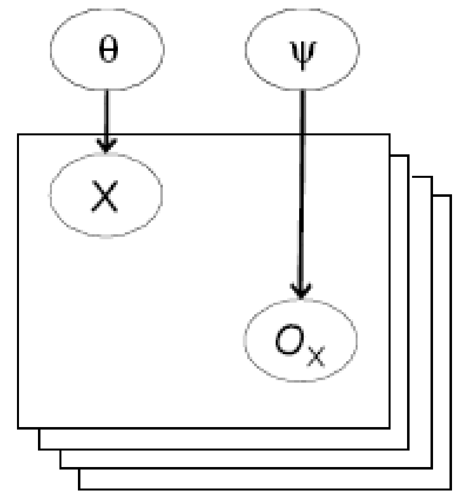
  $P_{miss}(\mathbf{O}_\mathbf{X} \mid \mathbf{X})$ parameterized by ψ
- $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_n\}$ $Val(Y_i) = Val(X_i) \cup \{?\}$

$$Y_i = \begin{cases} X_i & \text{if } +o_i \\ ? & \text{if } -o_i \end{cases}$$

Uncorrelated Missingness



Thumbtack falls off table, ...not recorded



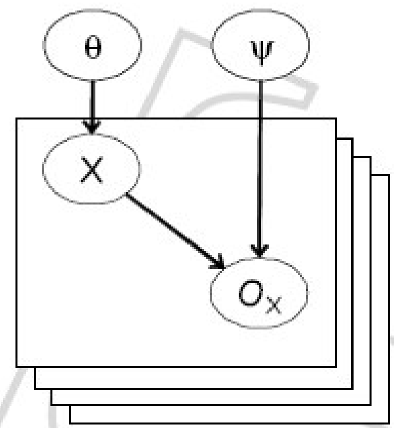
- Here, $\mathbf{X} \perp \mathbf{O}_\mathbf{X}$; $\theta \perp \psi \mid D$
  - $P(\ Y = H\ ) = \theta\ \psi$
    $P(\ Y = T\ ) = (1 - \theta)\ \psi$
    $P(\ Y = ?\ ) = 1 - \psi$
- Assuming D contains $\#[H]$, $\#[T]$, $\#[?]$
  - $L[\theta\ ,\ \psi : D] = \theta^{\#[H]}\,(1 - \theta)^{\#[T]}\ \psi^{\#[H]+\#[T]}\ (1 - \psi\ )^{\#[?]}$
  - $\theta^{(MLE)} = \#[H] \,/\, (\#[H] + \#[T])$
  - $\psi^{(MLE)} = (\#[H] + \#[T]) \,/\, (\#[H] + \#[T] + \#[?])$

Simple frequencies!!

Correlated Missingness

Recorder doesn't like "Tails", and so omits those values

- Here, $\neg[\ \theta \perp \psi \mid D\ ]$
  - $\psi_{Ox|H}$ = prob of seeing output, if heads $= P(\ Y=H \mid X=H)$
    $\psi_{Ox|T} = P(\ Y=T \mid X=T)$



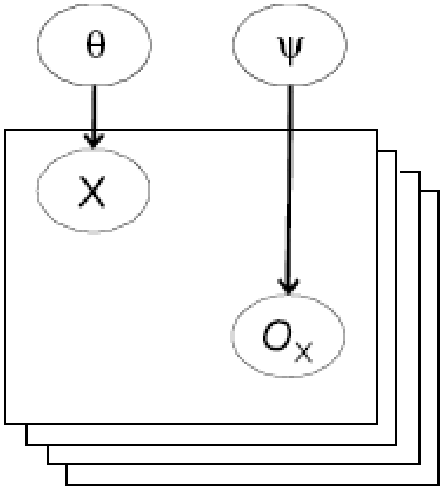
  - $P(\ Y=H\ ) = \theta\ \psi_{Ox|H}$
    $P(\ Y=T) = (1-\theta)\ \psi_{Ox|T}$
    $P(\ Y=?) = \theta\ (1-\psi_{Ox|H}) + (1-\theta)\ (1-\psi_{Ox|T})$
- Assuming D contains #[H], #[T], #[?]
  - $L[\theta\ ,\psi : D] = \theta^{\#[H]}\ (1-\theta)^{\#[T]}\ \psi_{Ox|H}{}^{\#[H]}\ \psi_{Ox|T}{}^{\#[T]}$
    $(\theta\ (1-\psi_{Ox|H}) + (1-\theta)\ (1-\psi_{Ox|T})\ )^{\#[?]}$

What a mess! Does not factor, so no easy MLE values…

Missing Completely At Random

A missing data model $P_{missing}$ is *missing completely at random (MCAR)* if

$P \models \mathbf{X} \perp \mathbf{O}_{\mathbf{X}}$



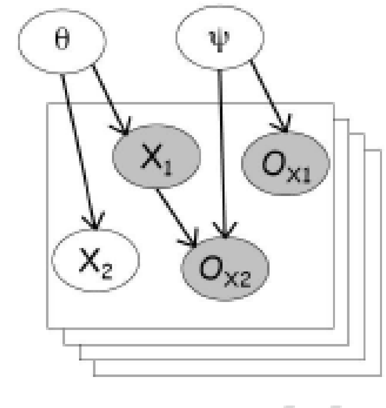
- Plausible …
  - Coffee spills on paper
  - Flecks of dusts in images
- Here, can solve separately for
  - θ (for $P(\mathbf{X})$)
  - ψ (for $P_{miss}(\mathbf{O}_{\mathbf{X}} \mid \mathbf{X})$)

MCAR is … too strong !

- Not MCAR:
  - test results are missing if not ordered…
    perhaps as patient too sick or too healthy
  - $\Rightarrow$ Missingness-of-test is correlated with test-outcome
- MCAR is sufficient for decomposition of likelihood…
  - but NOT necessary
- Just need

Observation mechanism is
CONDITIONALLY INDEPENDENT of variables,
GIVEN OTHER OBSERVATIONS

Weaker Condition



- Flip coin X1, X2
- If X1=Heads, reveal outcome of X2

- Here, $P \models O_{X2} \perp X_2 \mid X_1$
  - Outcomes of both coins INDEPENDENT of whether hidden, given observations
- Use θ_{X1} θ_{X2} $\psi_{Ox2|H}$ $\psi_{Ox2|T}$ (where $\theta_{X1} \perp \theta_{X2}$)

$$\begin{aligned} L(\theta : \mathcal{D}) \;=\; & \theta_{X_1}^{M[Y_1=Heads]}(1-\theta_{X_1})^{M[Y_1=Tails]} \\ & \theta_{X^2}^{M[Y_2=Heads]}(1-\theta_{X_2})^{M[Y_2 Tails]} \\ & \psi_{O_{X_2}|H}^{M[Y_1=Heads,Y_2=Heads]+M[Y_1=Heads,Y_2=Tails]}(1-\psi_{O_{X_2}|H})^{M[Y_1=Heads,Y_2=?]} \\ & \psi_{O_{X_2}|Tails}^{M[Y_1=Tails,Y_2=Heads]+M[Y_1=Tails,Y_2=Tails]}(1-\psi_{O_{X_2}|Tails})^{M[Y_1=Tails,Y_2=?]} \end{aligned}$$

- Four factors, each w/ just 1 parameter

  $\Rightarrow$ can solve independently!

Missing At Random

- Given tuple of observations y, partition variables $\mathbf{X}$ into
  - observed $X^y_{obs} = \{ X_i \mid y_i \neq ? \}$
  - hidden $X^y_{hid} = \{ X_i \mid y_i = ? \}$
- Missing data model P_{miss} is *missing at random (MAR)* if

  $\forall\ y$ w/ $P_{miss}(y)>0$ and $\forall\ x^y_{hid} \in Val(X^y_{hid})$

  $$P_{miss} \models o_X \perp x^y_{hid} \mid x^y_{obs}$$

$$P_{miss}(x^y_{hid} \mid x^y_{obs}, o_X) = P_{miss}(x^y_{hid} \mid x^y_{obs})$$

Meaning of MAR…

- MAR $\Rightarrow$

  $P_{miss}(x^y_{hid} \mid x^y_{obs}, o_X) = P_{miss}(x^y_{hid} \mid x^y_{obs})$

  $\Rightarrow$

  $P_{miss}(y) = P_{miss}(o_X \mid x^y_{obs}) \; P(x^y_{obs})$

Depends on ψ

Depends on θ

If P_{miss} is MAR, then

$L(\theta, \psi; D) = L(\theta; D) \, L(\psi; D)$

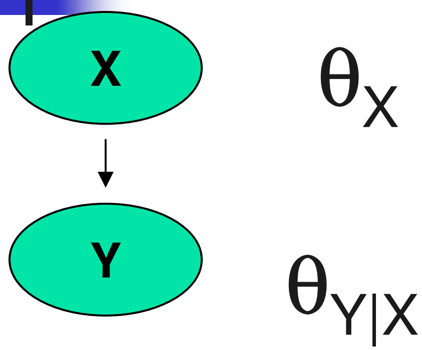
MAR $\Rightarrow$

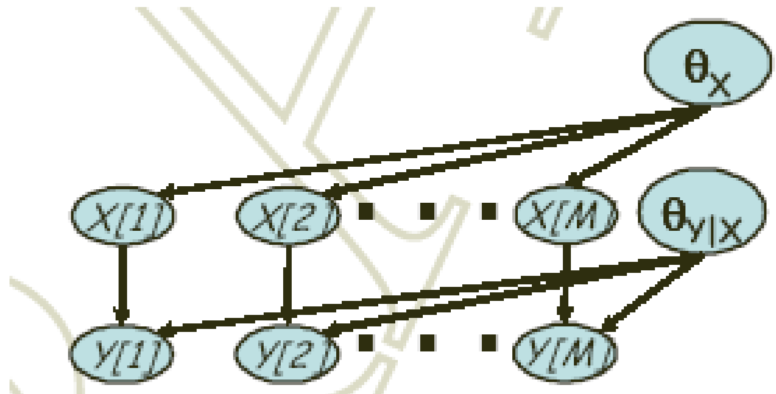
Can ignore observation model when learning model parameters!

Comments on MAR…

- There are many MAR situations but …
- BP_Sensor measures blood pressure
  - BP_Sensor can fail if patient is overweight
  - Obesity is relevant to blood pressure
  - So… "non-observation" is informative – not MAR
  - (But if we know Weight & Height, then $O_B \perp B \mid \{W,H\}$)
- Probably no X-ray X if no broken bones,
  - So $\neg(O_X \perp X)$, not MAR
  - But if "primary complaint" C known, $O_X \perp X \mid C$ … MAR!
- We will assume MAR from now on…

Bayesian Learning for 2-node BN





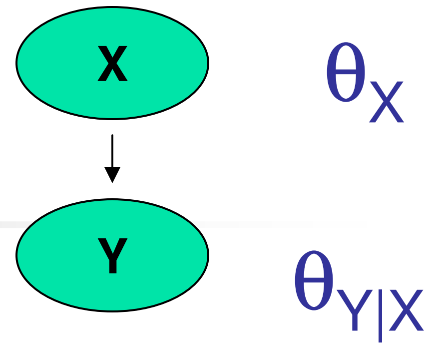
- Every path between $\theta_X - \theta_{Y|X}$ is:
  - $\theta_X \rightarrow X[m] \rightarrow Y[m] \leftarrow \theta_{Y|X}$
- ~~Complete~~ Partial data

  $\Rightarrow$ values for $\mathbf{D} = \{\cancel{X[1], \ldots, X[M],}\ Y[1], \ldots, Y[M]\}$

  $\Rightarrow$ path is ~~NOT~~ active

$\Rightarrow \quad \cancel{\theta_X \perp \theta_{Y|X} \mid \mathbf{D}}$

Example …

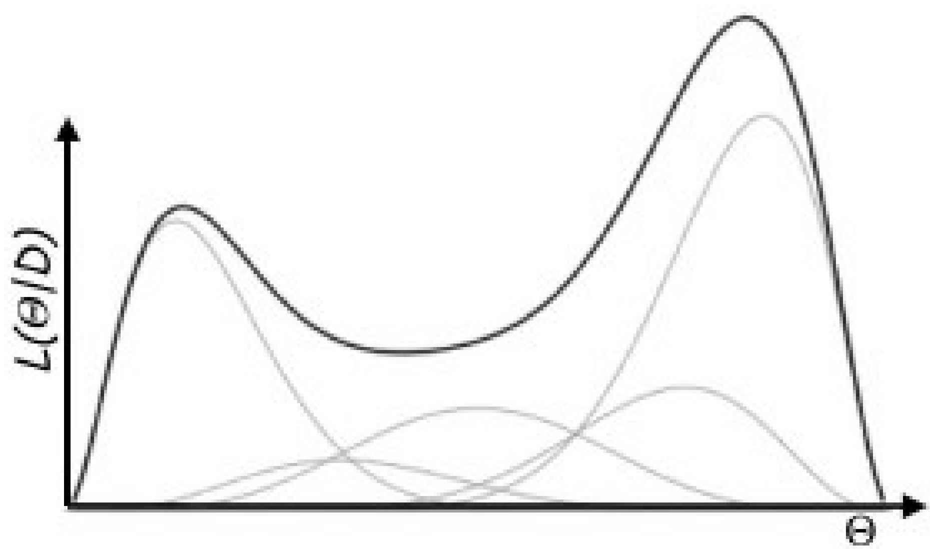


+x, +y:	13
+x, −y:	16
−x, +y:	10
−x, −y:	4

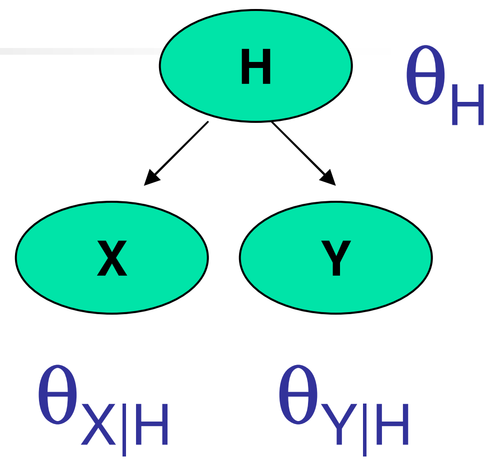
- Complete data:
- Likelihood:
  - $\theta_x^{29}(1-\theta_x)^{14}\,\theta_{y|+x}^{10}\,(1-\theta_{y|+x})^4\,\theta_{y|-x}^{13}\,(1-\theta_{y|-x})^{16}$
  - Easy to solve
- What if don’t know X[1]
  - (Assume Y[1]=+)
  - Likelihood:
    $\theta_x^{29}(1-\theta_x)^{13}\,\theta_{y|+x}^{10}\,(1-\theta_{y|+x})^4\,\theta_{y|-x}^{12}\,[\theta_x\,\theta_{y|+x}+(1-\theta_x)\,\theta_{y|-x}]$
  - Not as nice…
- If k missing values, L(…; D) could have many terms…

Geometric Visualization

- Complete data: *unimodal*
- Incomplete data:
  ... sum of unimodals...
  which is *multimodal* !



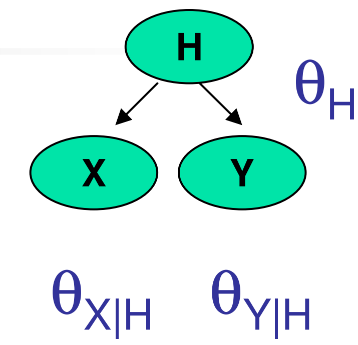
Problems with Hidden Variables



- Observe X, Y… but not H
  - $P(+x, -y) = \sum_h P(h)\, P(+x|h)\, P(-y|h)$
- Likelihood

  $L(\theta : D) = \prod_{x,y} [\sum_h P(h)\, P(x|h)\, P(y|h)\,]^{\#(x,y)}$
- Cannot decouple estimate of $P(x|h)$ from $P(y|h)$

Problems with Partial Data



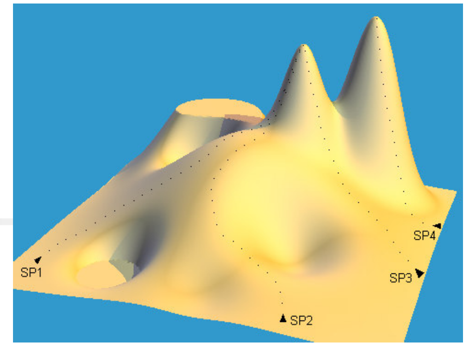
- In general, likelihood over iid data:
  $L(\theta : D) = \prod_m \left(\sum_{h[m]} P(\, o[m], h[m] \mid \theta)\right)$
- Involves *evaluating likelihood function* ...
  can be arbitrary BN inference $\Rightarrow$INTRACTABLE!

- More bad news: Likelihood function is…
  - *not* unimodal
  - does *not* have closed form representation
  - is *not* decomposable as product of likelihoods for diff parameters

Learning Belief Net Parameters from Partial Data

- Framework
  - Why is the data missing? ... MCAR, MAR, ...
  - Why more challenging?
- Approaches
  - Gradient Descent
  - EM
  - Gibbs

Gradient Ascent



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

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

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

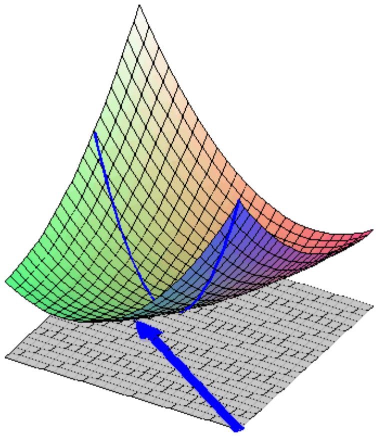
$$\frac{\partial \ln P_\Theta(S)}{\partial \theta_{ijk}} = \sum_{\ell=1}^{m} \frac{\partial \ln P_\Theta(c_\ell)}{\partial \theta_{ijk}} = \sum_{\ell=1}^{m} \frac{\partial P_\Theta(c_\ell)/\partial \theta_{ijk}}{P_\Theta(c_\ell)}$$

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

Alg: fn Basic-APN(BN = $\langle$ G, Θ $\rangle$, S): (modified) CPtables
    inputs: BN, a Belief net with CPT entries
        D, a set of data cases
  repeat until $\Delta\Theta \approx 0$
    $\Delta\Theta \leftarrow 0$
    for each $c_r \in S$
      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



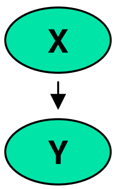
- Lots of Tricks for efficient ascent
  - Line Search
  - Conjugate Gradient
  - …

  Take Cmput551, or optimization

- 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 …

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

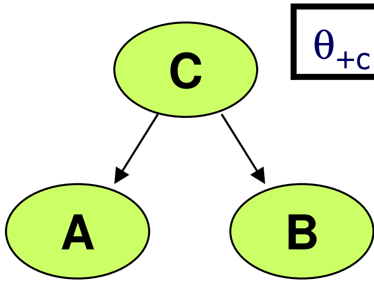


θ_X

$\theta_{Y|X}$

EM Approach – E Step

θ_{+c}	θ_{-c}



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

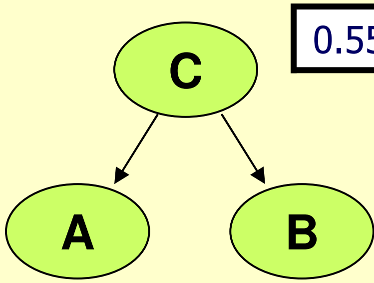
$\theta_{+b\|+c}$	$\theta_{-b\|+c}$
$\theta_{+b\|-c}$	$\theta_{-b\|-c}$

Sample S =

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

Guess initial values θ^0

0.55	0.45



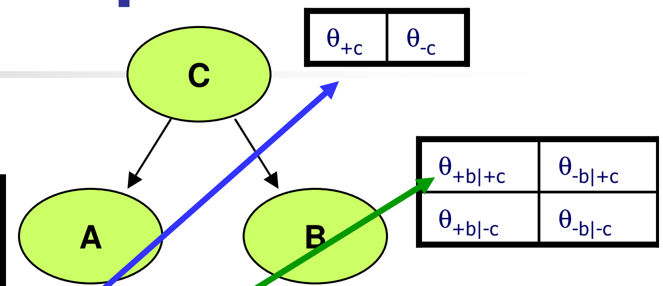
0.8	0.2
0.3	0.7

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:

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

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

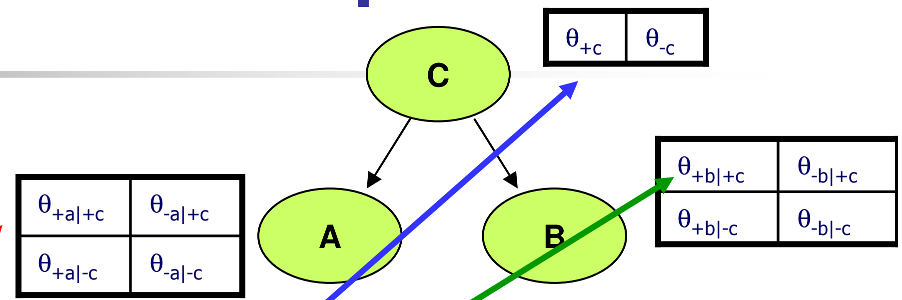
$$\hat{\theta}^{(1)}_{+c} = \frac{\#(+c)}{\#(\{\})} = \frac{1.0+(1.0)+(1.0)}{4} = 0.75$$

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:

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

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

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

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 θ,
  - find probability of each missing value
    - ... so get $E[\ N_{ijk}\ ]$
- **M step**:
  - Given completed (fractional) data
    - based on $E[\ N_{ijk}\ ]$
  - find max-likely parameters θ

EM Approach

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

- Iteratively, $k = 0, \ldots$

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

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

  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 …

- Always converges
- Always improve likelihood
  - $L(\theta^{(t+1)} : D) > L(\theta^{(t)} : D)$
  - … 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 $d_{ij}^{(0)} =$ Random[Domain(X_i)]

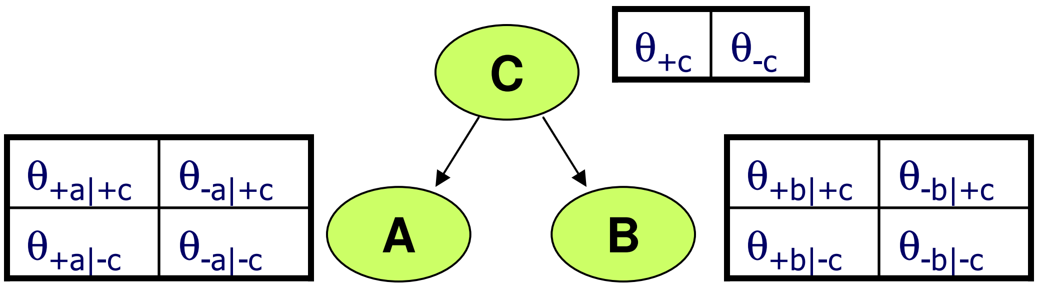
- For $k = 0..$
  - Compute $\Theta^{(k)}$ from $S^{(k)}$ [frequencies]
  - Form $S^{(k+1)}$ by...
    - $d_{ij}^{k+1} = c_{ij}$
    - If $c_{ij} = *$ then
      Let d_{ij}^{k+1} be random value for X_i,
      based on current distr Θ^k over $Z - X_i$

- 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

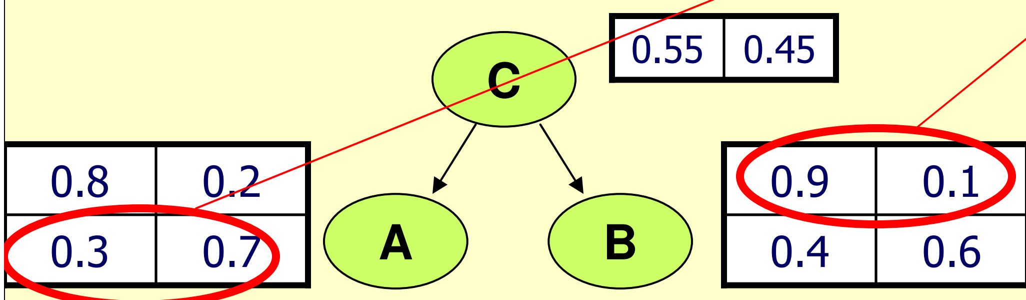


New $S^{(1)}$ =

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



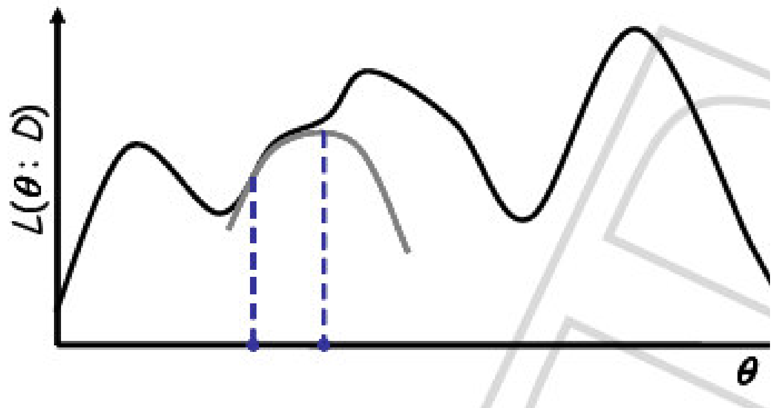
Then
- Use $S^{(1)}$ to get new $\theta^{(2)}$ parameters
- Form new $S^{(2)}$ by drawing new values from $\theta^{(2)}$

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 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…

Gaussian Approximation

(Assumes large amounts of data)

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

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

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

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

  ... which looks (approximately) Gaussian!

- Now use
  gradient descent or EM

Note: Can often use values computed during inference!

Summary

- Missingness: MCAR vs MAR
- Approaches
  - Gradient Ascent
  - EM
  - Gibbs sampling
    - Multiple imputation

- Note covered: Bayesian methods
  - MCMC, Variational, Particles, ...