Readings:
Murphy 11
(opt: K&F 18.6, 19.1, 19.2, 19.4)

Learning with Partially Observed Data

Prof. Suchi Saria

Slides adapted from versions by Profs. Eran Segal and Suin Lee

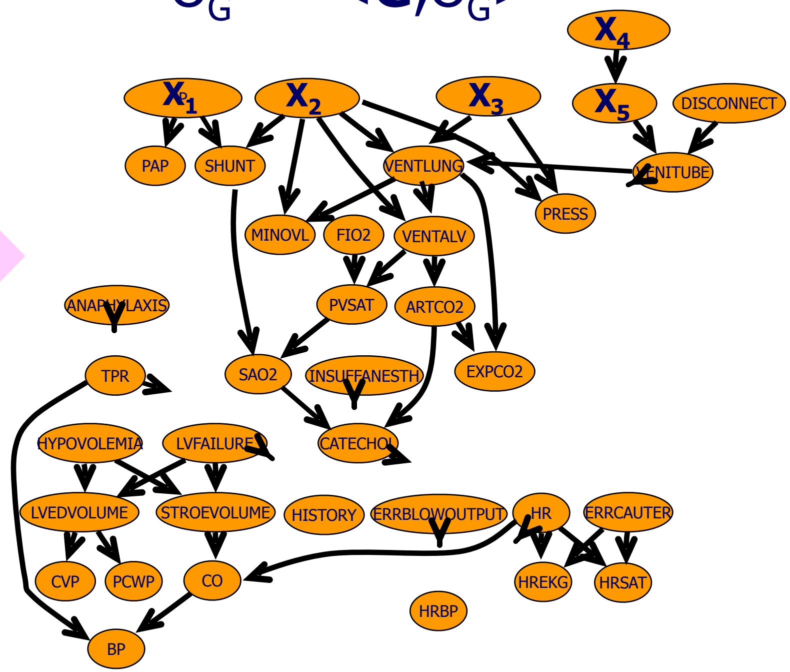
Training Data D

Training instance

D

X_1	3 1 0 -1 1 1 0 0 2 9 8 …
X_2	1 1 1 1 0 0 7 2 3 6 5 …
X_3	0 0 1 0 1 0 8 2 1 2 3 …
X_4	1 2 5 -2 3 0 1 3 4 5 …
:	:
X_{N-1}	1 3 2 3 6 …
X_N	0 7 4 -4 7 …

Θ_G $\quad$ $<\mathbf{G},\Theta_G>$



- Until now, we assumed that the training data is fully observed
  - Each instance assigns values to all the variables in our domain

Incomplete Data

- In reality, this assumption might not be true.

Training instance

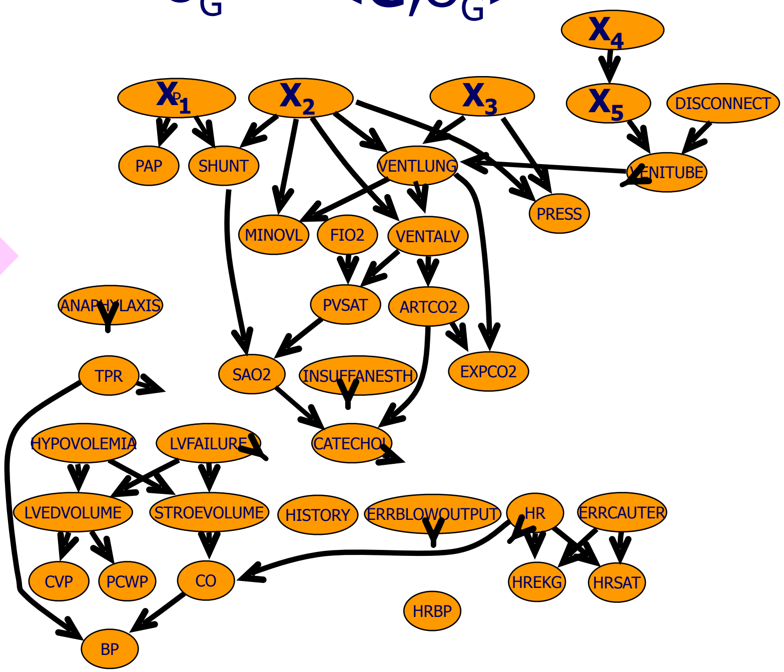
D

X_1 3 1 **?** -1 1 1 **?** 0 2 9 8 …
X_2 1 **?** 1 1 0 0 7 2 3 6 5 …
X_3 0 0 1 **?** 1 0 8 2 **?** 2 3 …
X_4 1 2 5 -2 **?** 0 1 3 4 5 …

Lung cancer?

? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?
X_{N-1} 0 7 4 **?** 7
X_N

Θ_G $<\mathbf{G},\Theta_G>$



- Missing values, Hidden variables

- Challenges
  - Foundational – is the learning task well defined?
  - Computational – how can we learn with missing data?

Types of Missingness

- Missing Completely at Random (MCAR): if the events that lead to the data-item being missing are independent both of observable variables and of unobservable parameters of interest, and occur entirely at random
    - WBC data has only been transcribed for a fraction of the patients by the administrator.
    - (Little et al., JASA 1988)
- Missing Not at Random (MNAR): is data that is missing for a specific reason (i.e. the value of the variable that's missing is related to the reason it's missing)
    - WBC is only measured when patient is sick. Sickness status not observed.
- Missing at Random (MAR): occurs when the missingness is related to a particular variable, but it is not related to the value of the variable that has missing data
    - Male participants are more likely to refuse to fill depression survey but willingness to fill does not depend on depression result.
    - REFS: Jaeger, ECML 2006; Tian, AISTATS 2015

More formally...

Need to understand what led to missing values

- **Missing Completely at Random (MCAR)**: Missingness is totally random; does not depend on anything.

  R models whether an element is missing or not. Y is the quantity of interest. (E.g., say your goal were to estimate the mean of Y.)

  – $P(R|Y,X) = P(R|Y,X^{obs},X^{mis}) = P(R|\psi)$

  – Cases with missing values a random sample of the original sample

  – No systematic differences between those with missing and observed values

  – Analyses using only complete cases will not be biased, but may have low power

  – Generally unrealistic, although may be reasonable for things like data entry errors

- **Missing At Random (MAR)**: Missingness depends on observed data
  - $P(R|Y,X) = P(R|Y,X^{obs},\psi)$
  - e.g., women more likely to respond than men
  - So there are differences between those with observed and missing values, but we observe the ways in which they differ
  - Can use weighting or imputation approaches to deal with the missingness
  - This is probably the assumption made most frequently
  - Including a lot of predictors in the imputation model can make this more plausible

- **Not Missing At Random (NMAR)**: Missingness depends on unobserved values
  - $(R|Y, X)$ cannot be simplified
  - e.g., probability of someone reporting their income depends on what their income is
  - e.g., probability of reporting prior arrests depends on whether or not they had previously been arrested
  - e.g., probability of reporting prior arrests depends on whether or not they are left-handed, and we do not observe left-handedness for anyone
  - i.e., even among people with the same values of the observed covariates, those with missing values on Y have a different distribution of Y than do those with observed Y
  - So we can't just use the observed cases to help impute the missing cases
  - Unfortunately no easy ways of dealing with this...have to posit some model of the missing data process

Treating Missing Data

- How should we treat missing data?
  - Based on data missing mechanism

- Case I: A coin is tossed on a table, occasionally it drops and measurements are not taken (random missing)
  - Sample sequence: H,T,?,?,T,?,H
  - Treat missing data by ignoring it

- Case II: A coin is tossed, but only heads are reported (deliberate missing values)
  - Sample sequence: H,?,?,?,H,?,H
  - Treat missing data by filling it with Tails

⬇

We need to consider the data missing mechanism

Modeling Data Missing Mechanism

- Let's try to model the data missing mechanism

- $X = \{X_1,\ldots,X_n\}$ are random variables

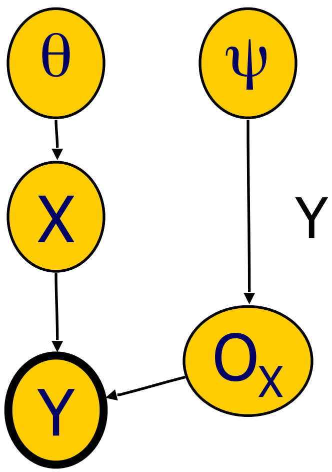
- $O_X = \{O_{X_1},\ldots,O_{X_n}\}$ are observability variables
  - Always observed

- $Y = \{Y_1,\ldots,Y_n\}$ new random variables
  - $Val(Y_i) = Val(X_i) \cup \{?\}$
  - Y_i is a deterministic function of X_i and O_{X_1}:

$$Y_i = \begin{cases} X_i & O_{X_i} = o^1 \\ ? & O_{X_i} = o^0 \end{cases}$$

Modeling Missing Data Mechanism

Case I
(random missing values)



Y=X if X is observed

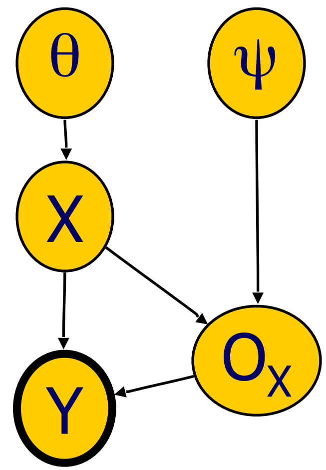
$$P(Y=H) = \theta\psi$$

$$P(Y=T) = (1-\theta)\psi$$

$$P(Y=?) = (1-\psi)$$

$$L(D:\theta,\psi) = \theta^{M_H} \cdot (1-\theta)^{M_T} \cdot \psi^{M_H+M_T} \cdot (1-\psi)^{M_?}$$

Case II
(deliberate missing values)



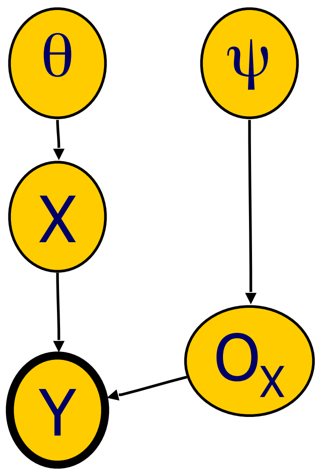
MLE

$$\hat{\theta} = \frac{M_H}{M_H + M_T}$$

$$\hat{\psi} = \frac{M_H + M_T}{M_H + M_T + M_?}$$

Modeling Missing Data Mechanism

Case I
(random missing values)



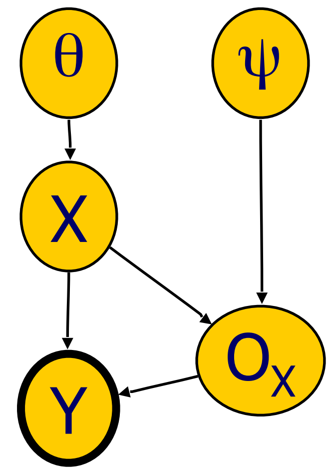
MLE ?

$\hat{\theta} =$

?

$\hat{\psi} =$

Case II
(deliberate missing values)



$$P(Y = H) = \theta\psi_{O_X|H}$$

$$P(Y = T) = (1-\theta)\psi_{O_X|T}$$

$$P(Y = ?) = \theta(1-\psi_{O_X|H}) + (1-\theta)(1-\psi_{O_X|T})$$

$$L(D:\theta,\psi) = \theta^{M_H} \cdot (1-\theta)^{M_T} \cdot \psi_{O_X|H}^{M_H} \cdot \psi_{O_X|T}^{M_T} \cdot \left(\theta(1-\psi_{O_X|H}) + (1-\theta(1-\psi_{O_X|T})\right)^{M_?}$$

- In the MAR and MCAR case, the true joint is recoverable.
  - Missing Completely at Random (MCAR)
    - For every X_i, $Ind(X_i;O_{X_i})$, a very strong assumption
    - Sufficient but not necessary for the decomposition of the likelihood
  - Missing at Random (MAR) is sufficient
    - The probability that the value of X_i is missing is independent of its actual value, given other observed values

- For a more general criteria, see paper by Shiptser et al., 2015 (posted in Resources)

Incomplete Data

- In reality, this assumption might not be true.

Θ_G $\langle G, \Theta_G \rangle$

D

X_1 3 1 ? -1 1 1 ? 0 2 9 8 …

X_2 1 ? 1 1 0 0 7 2 3 6 5 …

X_3 0 0 1 ? 1 0 8 2 ? 2 3 …

X_4 1 2 5 -2 ? 0 1 3 4 5 …

Lung cancer?

? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?

X_{N-1} 0 7 4 ? 7

X_N

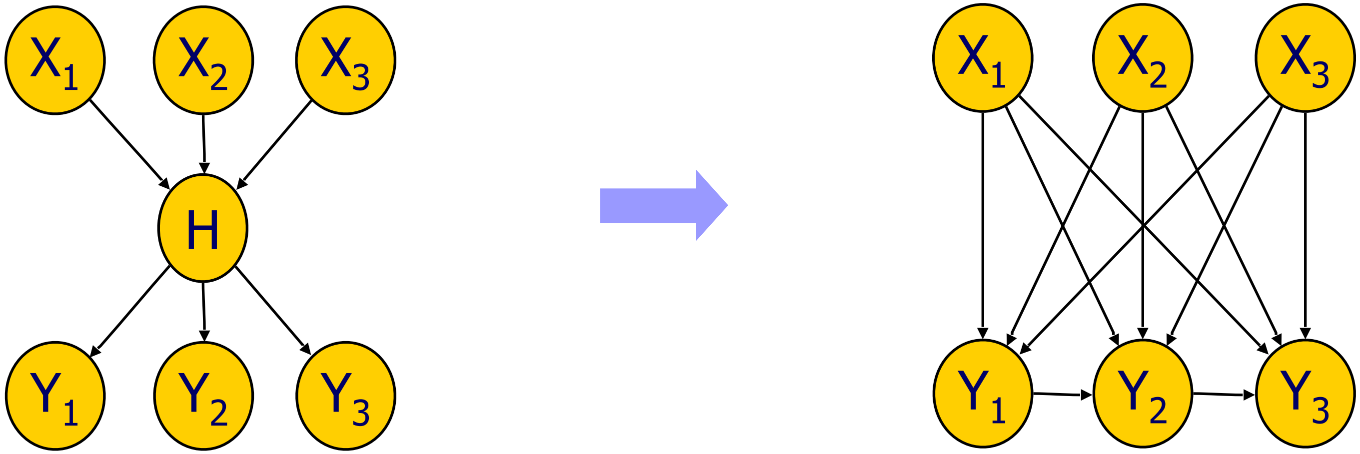
- Missing values, **Hidden variables**

- Challenges
  - Foundational – is the learning task well defined?
  - Computational – how can we learn with missing data?

Hidden (Latent) Variables

- Attempt to learn a model with hidden variables
  - In this case, MCAR always holds (variable is always missing)

- Why should we care about unobserved variables?



17 parameters

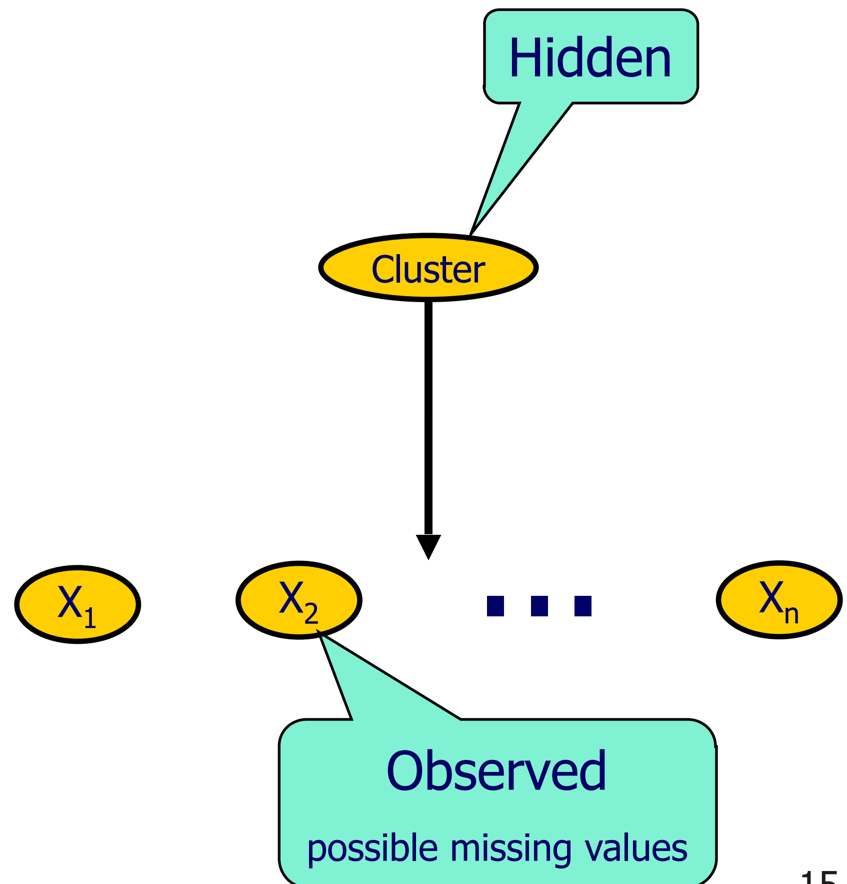
59 parameters

Hidden (Latent) Variables

- Hidden variables also appear in **clustering**
- Naïve Bayes model:
  - Class variable is hidden
  - Observed attributes are independent given the class

D

X_1	3 1 0 -1 1 1 0 0 2 9 8 …
X_2	1 1 1 1 0 0 7 2 3 6 5 …
X_3	0 0 1 0 1 0 8 2 1 2 3 …
X_4	1 2 5 -2 3 0 1 3 4 5 …
:	:
X_{N-1}	1 3 2 3 6 …
X_N	0 7 4 -4 7 …
H	1 2 1 1 3 3 1 1 2 2 1 1 …



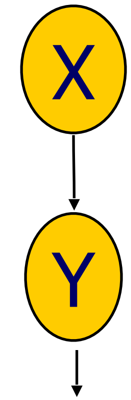
How do missing data affect the likelihood function?

Likelihood for Complete Data

Input Data:

X	Y
x^0	y^0
x^0	y^1
x^1	y^0

P(X)	
x^0	x^1
θ_{x0}	θ_{x1}



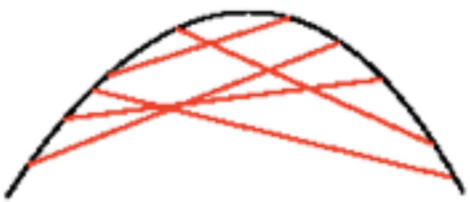
	P(Y|X)	
X	y^0	y^1
x^0	$\theta_{y0\|x0}$	$\theta_{y1\|x0}$
x^1	$\theta_{y0\|x1}$	$\theta_{y1\|x1}$

Likelihood:

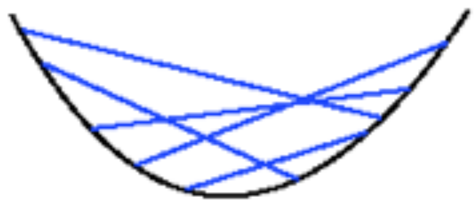
$$
\begin{aligned}
L(D:\theta) &= P(x[1],y[1])\cdot P(x[2],y[2])\cdot P(x[3],y[3]) \\
&= P(x^0,y^0)\cdot P(x^0,y^1)\cdot P(x^1,y^0) \\
&= \theta_{x^0}\cdot\theta_{y^0|x^0}\cdot\theta_{x^0}\cdot\theta_{y^1|x^0}\cdot\theta_{x^1}\cdot\theta_{y^0|x^1} \\
&= \left(\theta_{x^0}\cdot\theta_{x^0}\cdot\theta_{x^1}\right)\cdot\left(\theta_{y^0|x^0}\cdot\theta_{y^1|x^0}\right)\left(\theta_{y^0|x^1}\right)
\end{aligned}
$$

- Likelihood decomposes by variables
- Likelihood decomposes within CPDs
- Likelihood function is log-concave → unique global maximum that has a simple analytic closed form.

Example of concave, convex functions in R



A concave function.
No line segment lies above the graph at any point.



A convex function.
No line segment lies below the graph at any point.

A function that is neither concave nor convex.
The line segment shown lies above the graph at some points and below it at other points.

Theorem. Let f be a convex function, and let X be a random variable. Then:

$$\mathrm{E}[f(X)] \geq f(\mathrm{E}X).$$

Moreover, if f is strictly convex, then $\mathrm{E}[f(X)] = f(\mathrm{E}X)$ holds true if and only if $X = \mathrm{E}[X]$ with probability 1 (i.e., if X is a constant).

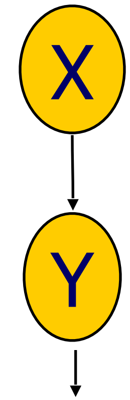
Source: https://www.economics.utoronto.ca/osborne/MathTutorial/CV1F.HTM

Likelihood for Incomplete Data

Input Data:

X	Y
?	y^0
x^0	y^1
?	y^0

P(X)	
x^0	x^1
θ_{x0}	θ_{x1}



X	P(Y|X) y^0	y^1
x^0	$\theta_{y0\|x0}$	$\theta_{y1\|x0}$
x^1	$\theta_{y0\|x1}$	$\theta_{y1\|x1}$

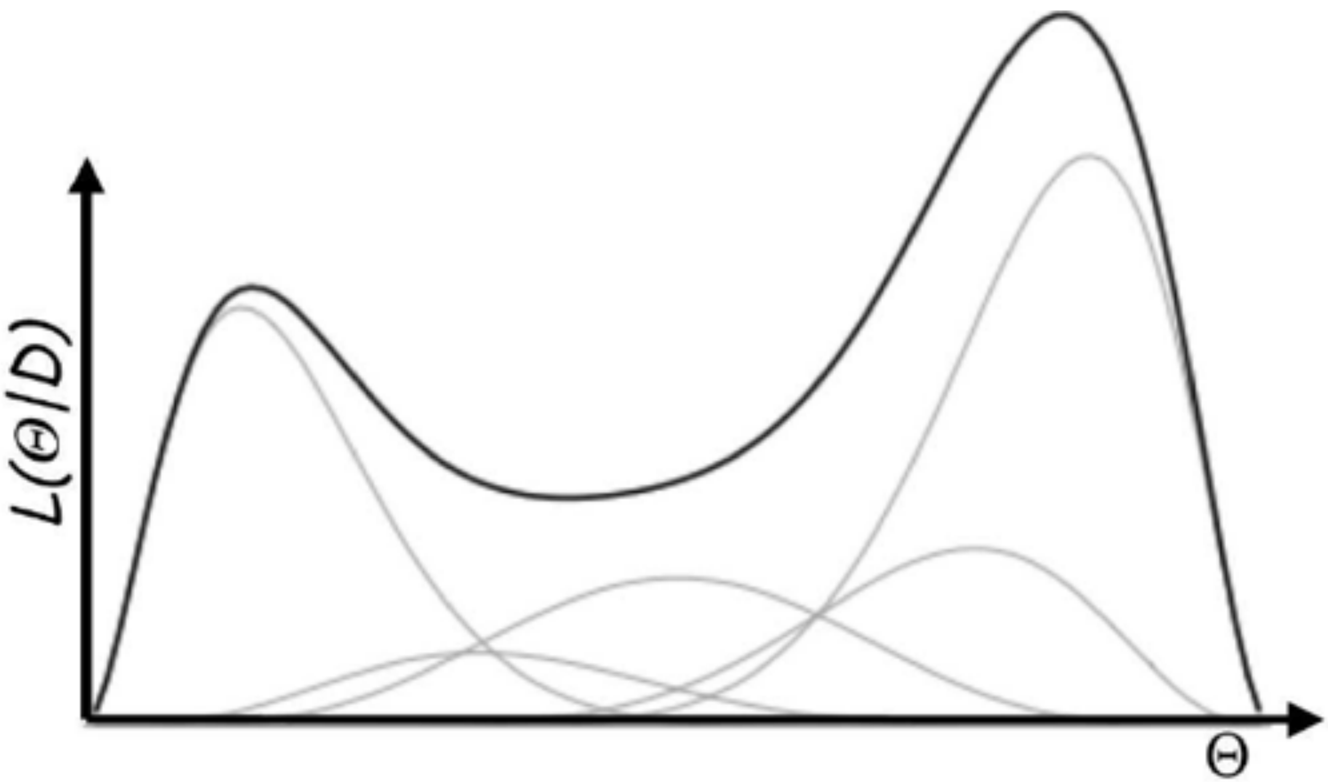
Likelihood:

$$
\begin{aligned}
L(D:\theta) &= P(y^0)\cdot P(x^0,y^1)\cdot P(y^0) \\
&= \left(\sum_{x\in X} P(x,y^0)\right)\cdot P(x^0,y^1)\cdot\left(\sum_{x\in X} P(x,y^0)\right) \\
&= \left(\theta_{x^0}\cdot\theta_{y^0|x^0}+\theta_{x^1}\cdot\theta_{y^0|x^1}\right)\theta_{x^0}\cdot\theta_{y^1|x^0}\cdot\left(\theta_{x^0}\cdot\theta_{y^0|x^0}+\theta_{x^1}\cdot\theta_{y^0|x^1}\right) \\
&== \left(\theta_{x^0}\cdot\theta_{y^0|x^0}+\theta_{x^1}\cdot\theta_{y^0|x^1}\right)^2\cdot\theta_{x^0}\cdot\theta_{y^1|x^0}
\end{aligned}
$$

- Likelihood does not decompose by variables
- Likelihood does not decompose within CPDs
- To be seen: Computing likelihood per instance requires inference!

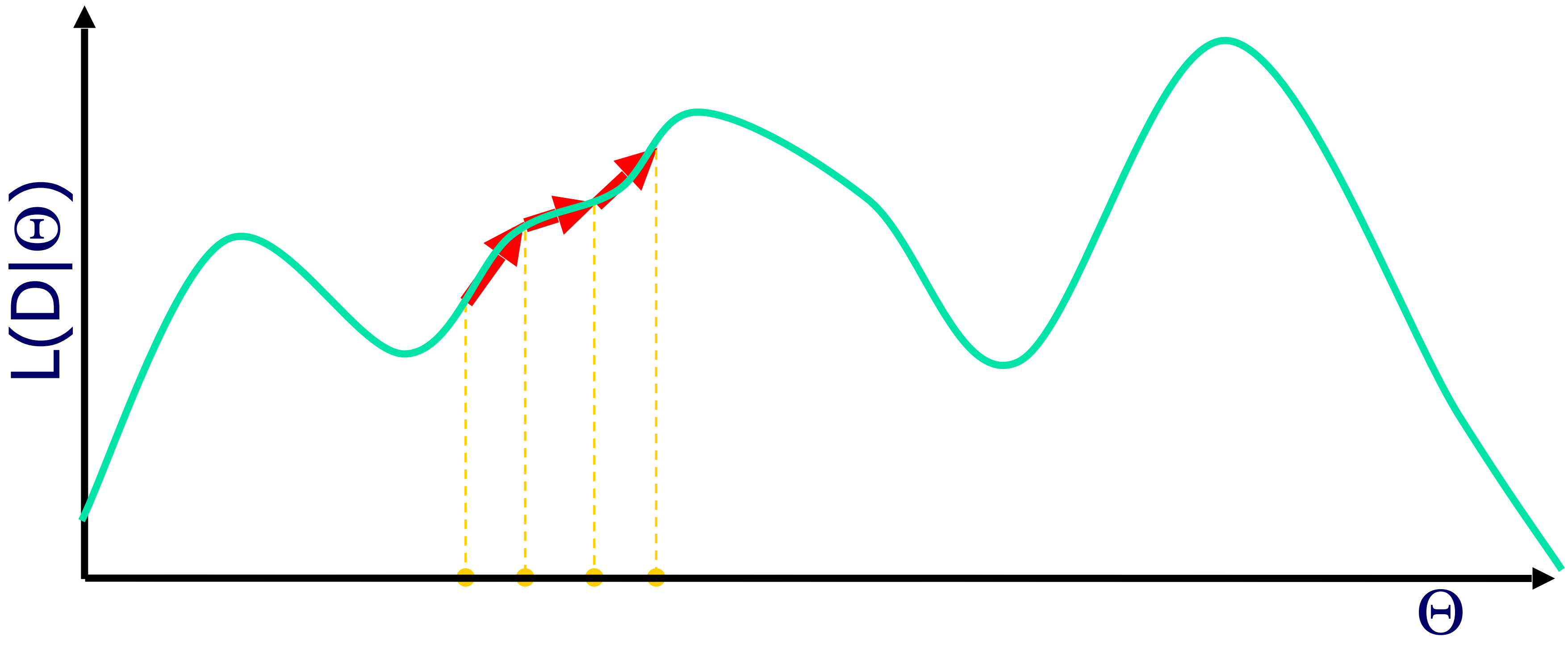
Likelihood with Missing Data

- Multimodal likelihood function with incomplete data
  - Likelihood function is not log-concave → local maxima cannot be obtained by a simple analytic closed form



MLE from Incomplete Data

- Take steps proportional to the positive of the gradient.

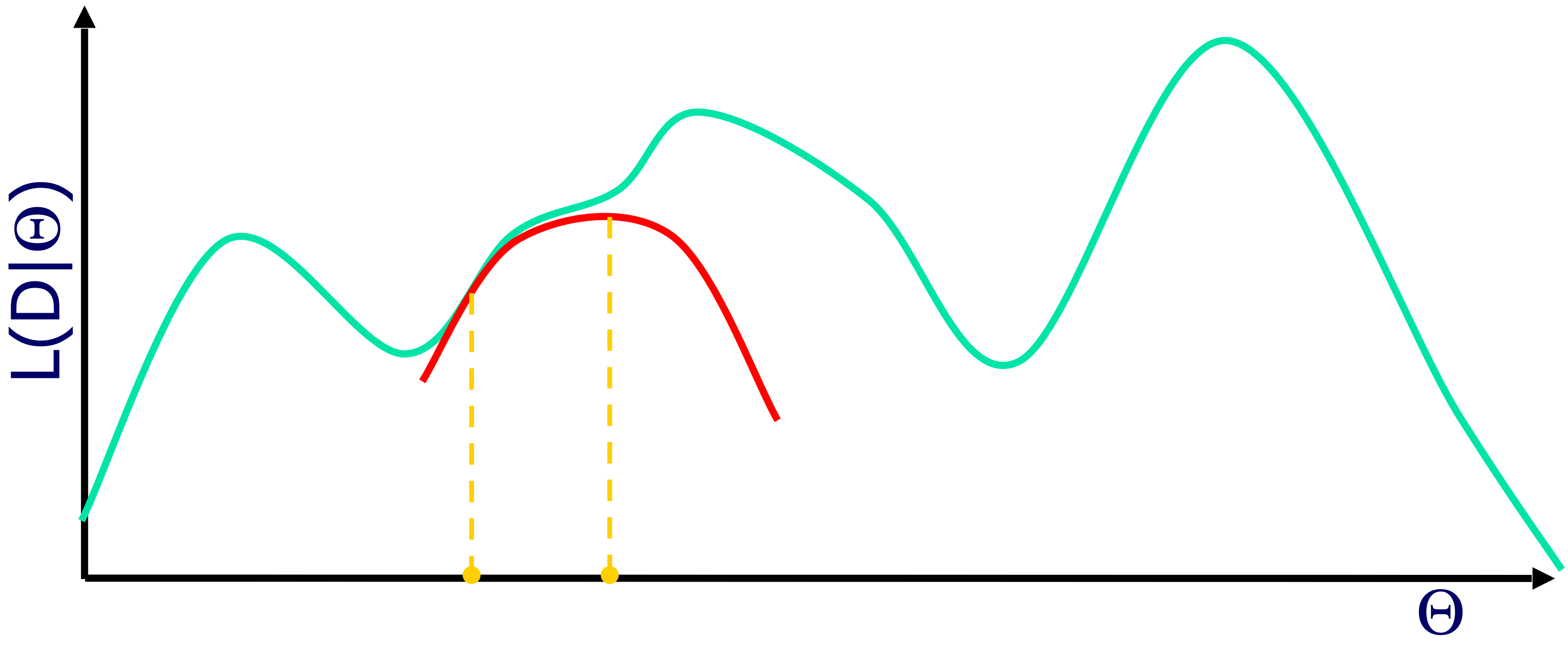


Gradient Ascent:

- Follow gradient of likelihood w.r.t. to parameters
- Add line search and conjugate gradient methods to get fast convergence

MLE from Incomplete Data

- Nonlinear optimization problem

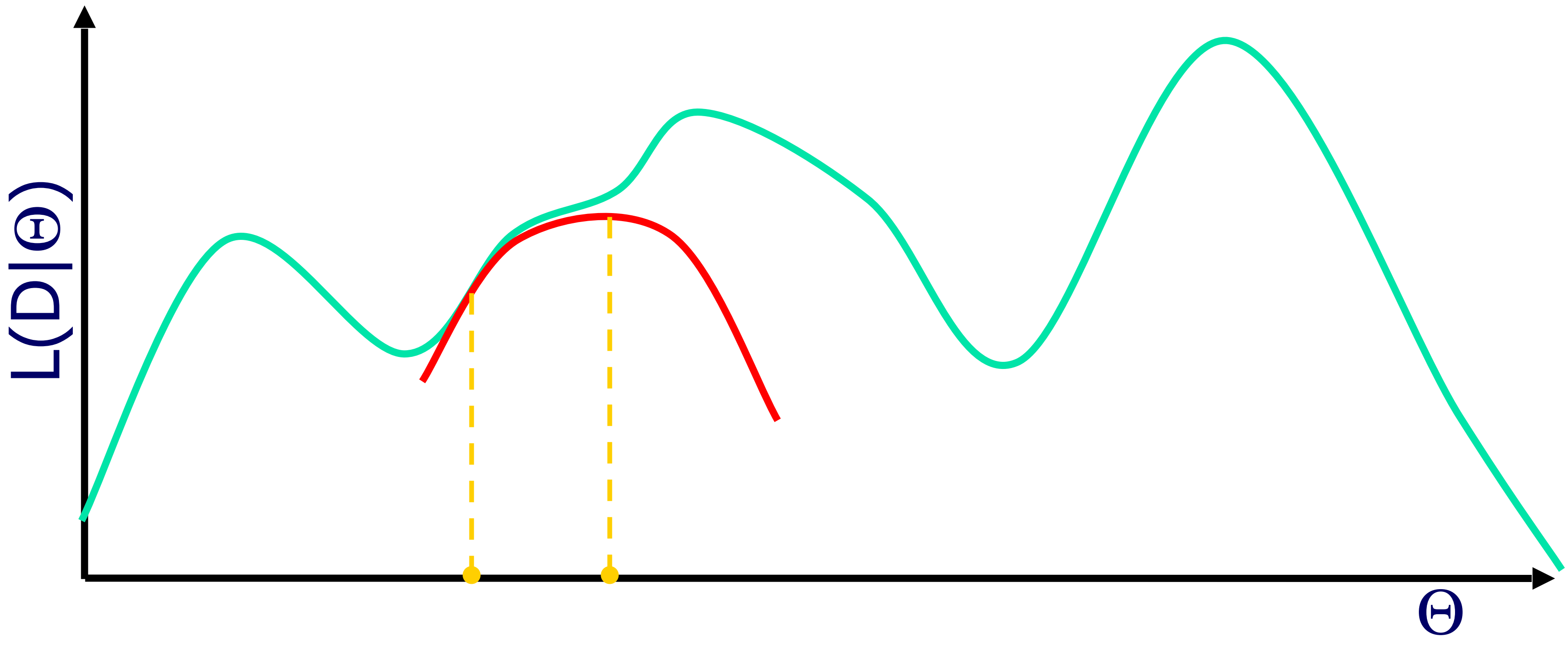


Expectation Maximization (EM):

- Use “current point” to construct alternative function (which is “nice”)
- Guaranty: maximum of new function has better score than current point

MLE from Incomplete Data

- Nonlinear optimization problem



Gradient Ascent and EM

- Find local maxima
- Require multiple restarts to find approx. to the global maximum
- Require computations in each iteration

Expectation Maximization (EM)

- Tailored algorithm for optimizing likelihood functions

- Intuition
  - Parameter estimation is easy given complete data
  - Computing probability of missing data is "easy" (=inference) given parameters

- Strategy
  - Pick a starting point for parameters
  - "Complete" the data using current parameters
  - Estimate parameters relative to data completion
  - Iterate
  - Procedure guaranteed to improve at each iteration

Deriving Expectation Maximization (EM)

$$\ell(\theta) = \sum_{i=1}^{m} \log p(x;\theta)$$

x is always observed
z is missing

$$\begin{aligned}\sum_i \log p(x^{(i)};\theta) &= \sum_i \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)};\theta) \\ &= \sum_i \log \sum_{z^{(i)}} Q_i(z^{(i)}) \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})} \\ &\geq \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}\end{aligned}$$

Jensen's inequality

If X is a random variable and ψ is a concave function, then

$$\psi(E[X]) \geq E[\psi(X)]$$

See http://cs229.stanford.edu/notes/cs229-notes8.pdf for a write up of this derivation.

Sufficient for the following to the case for the bound to be tight:

$$\frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} = c$$

$$Q_i(z^{(i)}) \propto p(x^{(i)}, z^{(i)}; \theta) \qquad \sum_z Q_i(z^{(i)}) = 1$$

$$\begin{aligned} Q_i(z^{(i)}) &= \frac{p(x^{(i)}, z^{(i)}; \theta)}{\sum_z p(x^{(i)}, z; \theta)} \\ &= \frac{p(x^{(i)}, z^{(i)}; \theta)}{p(x^{(i)}; \theta)} \\ &= p(z^{(i)} | x^{(i)}; \theta) \end{aligned}$$

Expectation Maximization (EM) Summarized

Repeat until convergence{

(E-step) For each i, set

$$Q_i(z^{(i)}) := p(z^{(i)}|x^{(i)};\theta).$$

(M-step) Set

$$\theta := \arg\max_\theta \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}.$$

}

Expectation Maximization (EM)

- Initialize parameters to θ^0
- Iterate E-step and M-step

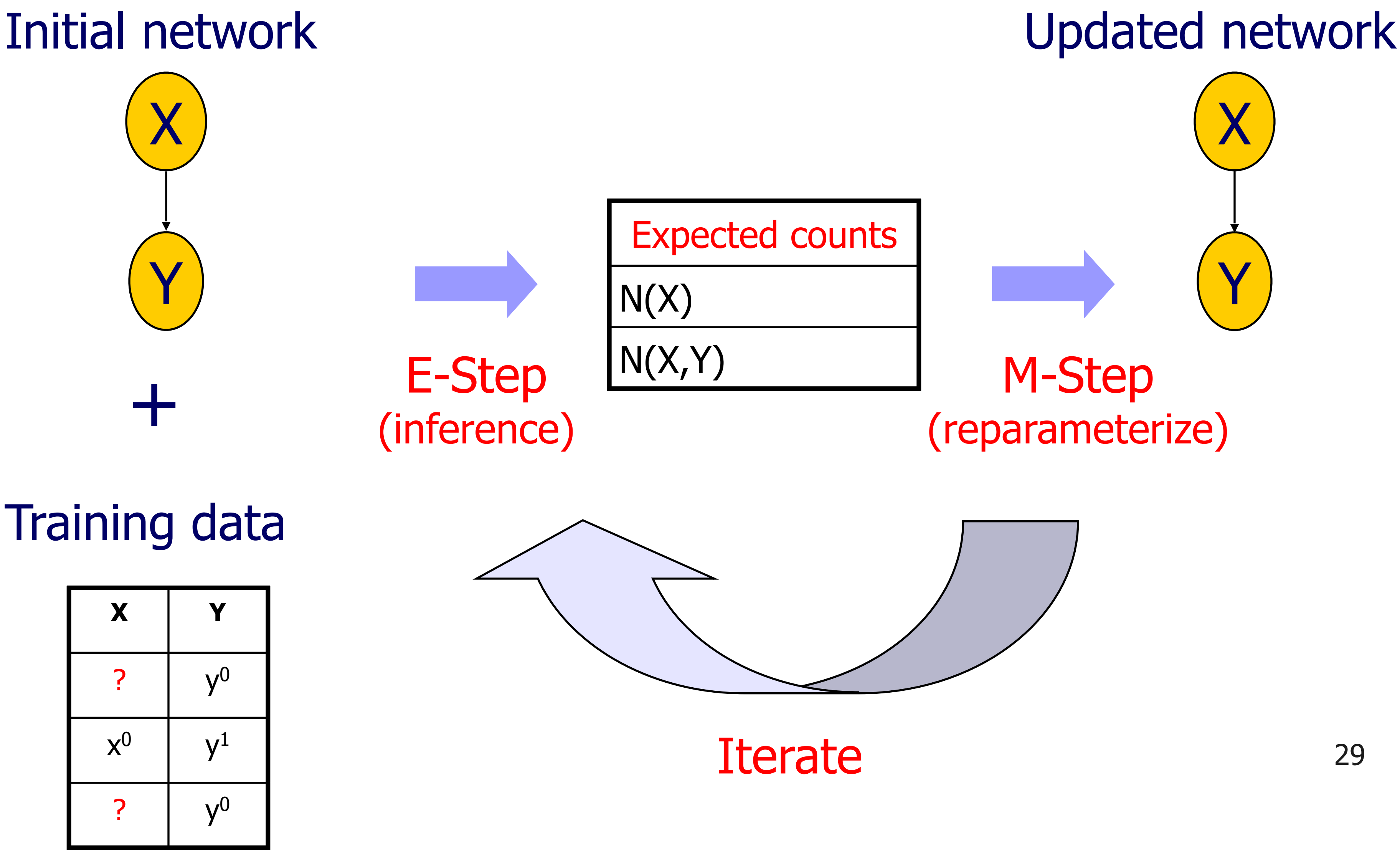
- In the t-th iteration, we do
- Expectation (E-step):
  - Let o[m] be the observed data in the m-th training instance.
  - For each m and each family X_i,$\mathbf{Pa_i}$, compute $P(X_i,\mathbf{Pa_i} \mid o[m], \theta^{(t)})$
  - Compute the expected sufficient statistics for each values x, **u** on X_i,$\mathbf{Pa_i}$, respectively.

$$\overline{M}_{\theta^{(t)}}[X_i = x, \mathbf{Pa}_i = \mathbf{u}] = \sum_m P(X_i = x, \mathbf{Pa}_i = \mathbf{u} \mid o[m], \theta^{(t)})$$

- Maximization (M-step):
  - Treat the expected sufficient statistics as observed and set the parameters to the MLE with respect to the ESS

$$\theta^{(t+1)}_{X_i = x \mid \mathbf{Pa}_i = \mathbf{u}} = \frac{\overline{M}_{\theta^{(t)}}[X_i = x, \mathbf{Pa}_i = \mathbf{u}]}{\overline{M}_{\theta^{(t)}}[\mathbf{Pa}_i = \mathbf{u}]}$$

Expectation Maximization (EM)



X	Y
?	y^0
x^0	y^1
?	y^0

Expectation Maximization (EM)

- Formal Guarantees:
  - $L(D:\Theta^{(t+1)}) \geq L(D:\Theta^{(t)})$
    - Each iteration improves the likelihood
  - If $\Theta^{(t+1)}=\Theta^{(t)}$, then $\Theta^{(t)}$ is a stationary point of $L(D:\Theta)$
    - Usually, this means a local maximum

- Main cost:
  - Computations of expected counts in E-Step
  - Requires inference for each instance in training set
    - Exactly the same as in gradient ascent!

- Reading material on EM
  - Please read Andrew Ng's lecture note

EM – Practical Considerations

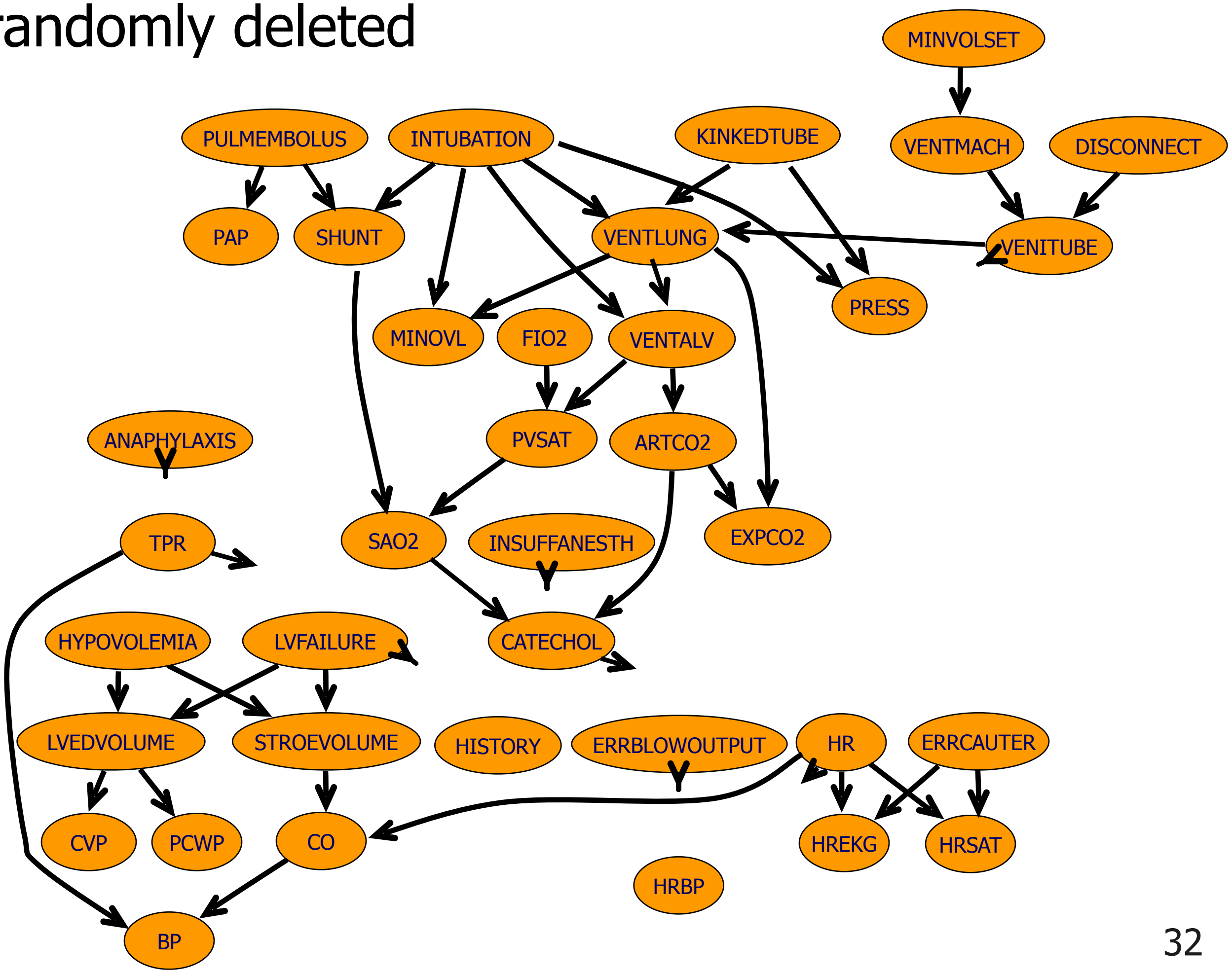
- Initial parameters
  - Highly sensitive to starting parameters
  - Choose randomly
  - Choose by guessing from another source

- Stopping criteria
  - Small change in data likelihood
  - Small change in parameters

- Avoiding bad local maxima
  - Multiple restarts
  - Early pruning of unpromising starting points

EM in Practice – Alarm Network

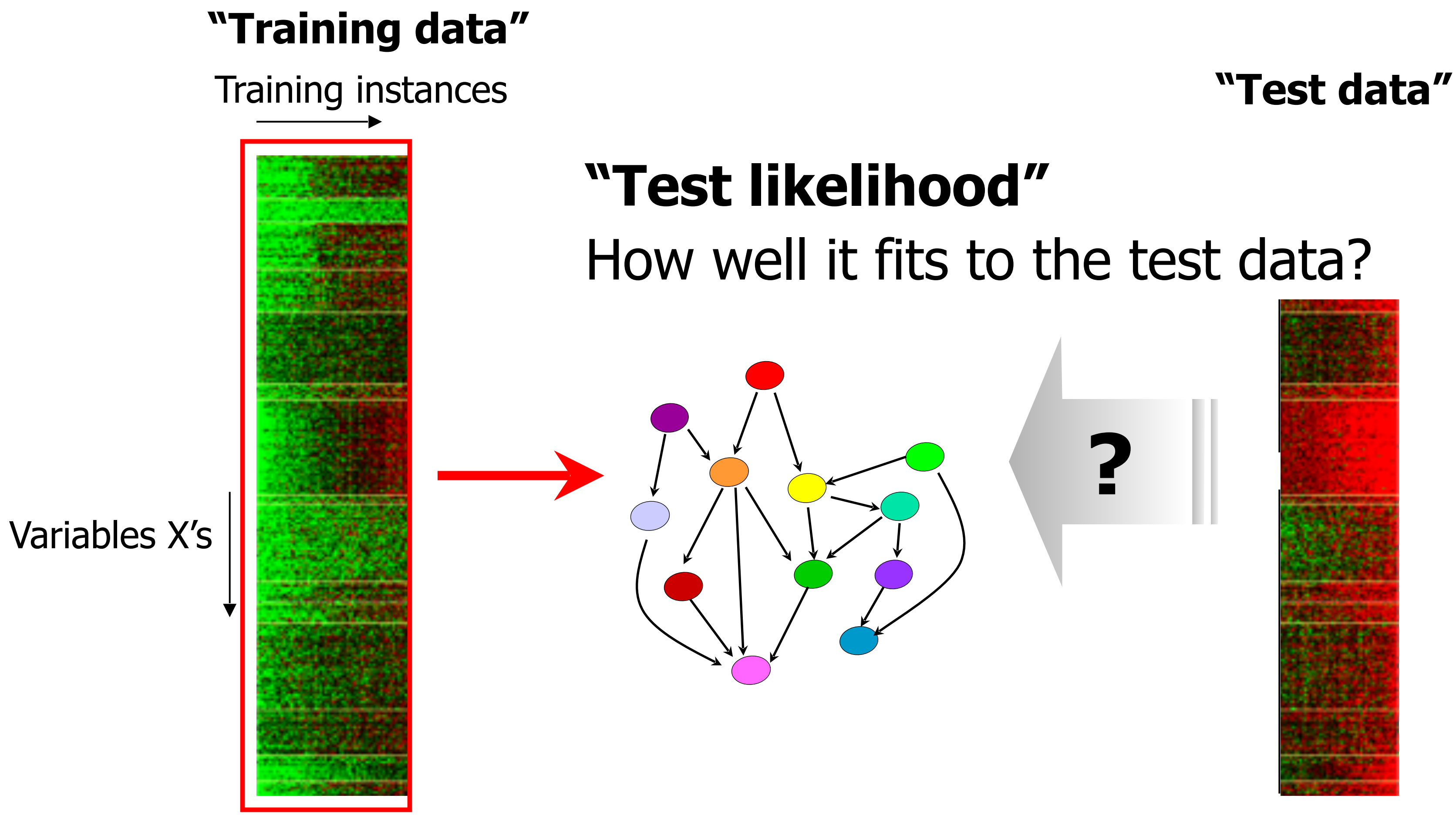
- Alarm network
  - Data sampled from true network
  - 20% of data randomly deleted



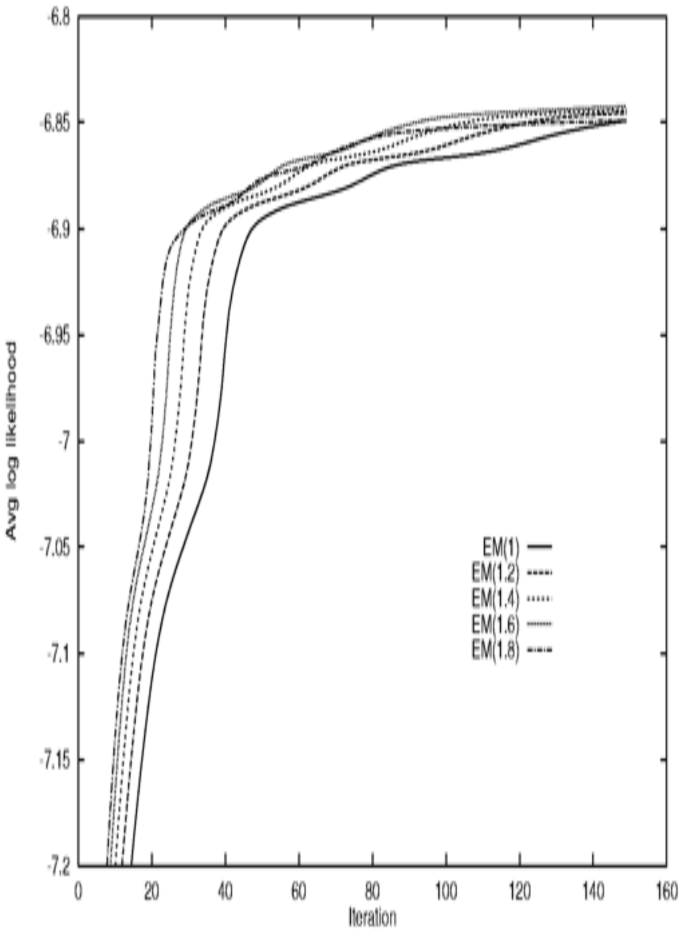
Update rules for parameter estimation in Bayesian networks
Eric Bauer, Daphne Koller, Yoram Singer, UAI 1997.

Statistical evaluation

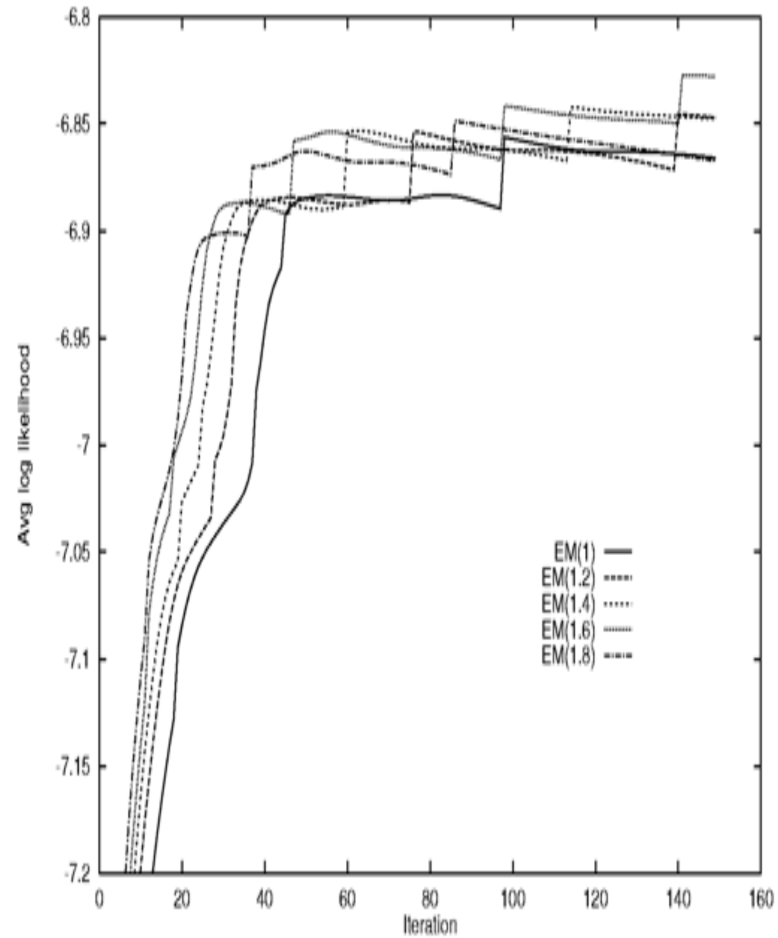
- Cross-validation test
  - Divide the data (experiments) into training and test data
  - Compute the likelihood function for the **Test data**



EM in Practice – Alarm Network



Training error



Test error

Partial Data: Parameter Estimation

- Non-linear optimization problem
- Methods for learning: EM and Gradient Ascent
  - Exploit inference for learning

- Challenges
  - Exploration of a complex likelihood/posterior
    - More missing data $\Rightarrow$ many more local maxima
    - Cannot represent posterior $\Rightarrow$ must resort to approximations
  - Inference
    - Main computational bottleneck for learning
    - Learning large networks $\Rightarrow$ exact inference is infeasible
      $\Rightarrow$ resort to approximate inference

Structure Learning w. Missing Data

- Distinguish two learning problems
  - Learning structure for a given set of random variables
  - Introduce new hidden variables
    - How do we recognize the need for a new variable?
    - Where do we introduce a newly added hidden variable within G?
    - Open ended and less understood…

Reading: Koller and Friedman 19.4

Structure Learning w. Missing Data

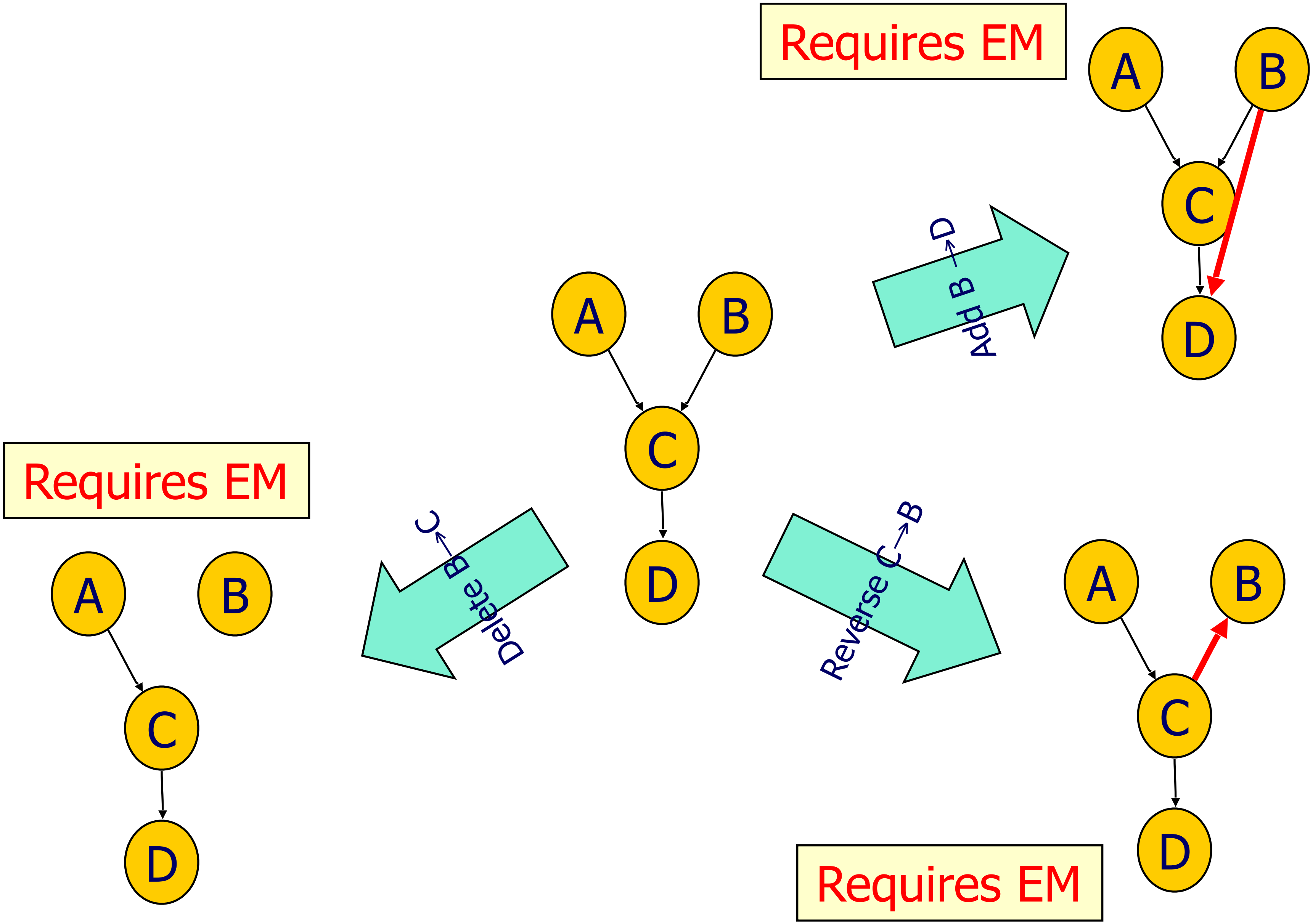
- Theoretically, there is no problem
  - Define score, and search for structure that maximizes it

$$Score_{BIC}(G : D) = l(\hat{\theta}_G : D) - \frac{\log M}{2} Dim(G)$$

  - Likelihood term will require gradient ascent or EM

- Practically infeasible
  - Typically we have $O(n^2)$ candidates at each search step
  - Requires EM for evaluating each candidate
  - Requires inference for each data instance of each candidate
  - Total running time per search step:
    $O(n^2 \cdot M \cdot \#EM\ iteration \cdot cost\ of\ BN\ inference)$

Typical Search



Structural EM

- Basic idea: use expected sufficient statistics to learn structure, not just parameters
  - Use current network to complete the data using EM
  - Treat the completed data as "real" to score candidates
  - Pick the candidate network with the best score
  - Use the previous completed counts to evaluate networks in the next step
  - After several steps, compute a new data completion from the current network

Structural EM

- Conceptually
  - Algorithm maintains an actual distribution Q over completed datasets as well as current structure G and parameters θ_G
  - At each step we do one of the following
    - Use $<G,\theta_G>$ to compute a new completion Q and redefine θ_G as the MLE relative to Q
    - Evaluate candidate successors G' relative to Q and pick best

- In practice
  - Maintain Q implicitly as a model $<G,\theta_G>$
  - Use the model to compute sufficient statistics $M_Q[x,\mathbf{u}]$ when these are needed to evaluate new structures
  - Use sufficient statistics to compute MLE estimates of candidate structures

Structural EM Benefits

- Many fewer EM runs
- Score relative to completed data is decomposable!

  - Utilize same benefits as structure learning w. complete data
  - Each candidate network requires few recomputations
  - Here savings is large since each sufficient statistics computation requires inference
- As in EM, we optimize a simpler score
- Can show improvements and convergence

Hint: See proof
Thm 19.5 KF

$$Score_{BIC}(\langle G,\theta_G\rangle : D) - Score_{BIC}(\langle G^Q,\theta_G^Q\rangle : D)$$
$$\geq E_Q\left[Score_{BIC}(\langle G,\theta_G\rangle : D^+)\right] - E_Q\left[Score_{BIC}(\langle G^Q,\theta_G^Q\rangle : D^+)\right]$$

  - An SEM step that improves in D^+ space, improves real score

Completed data