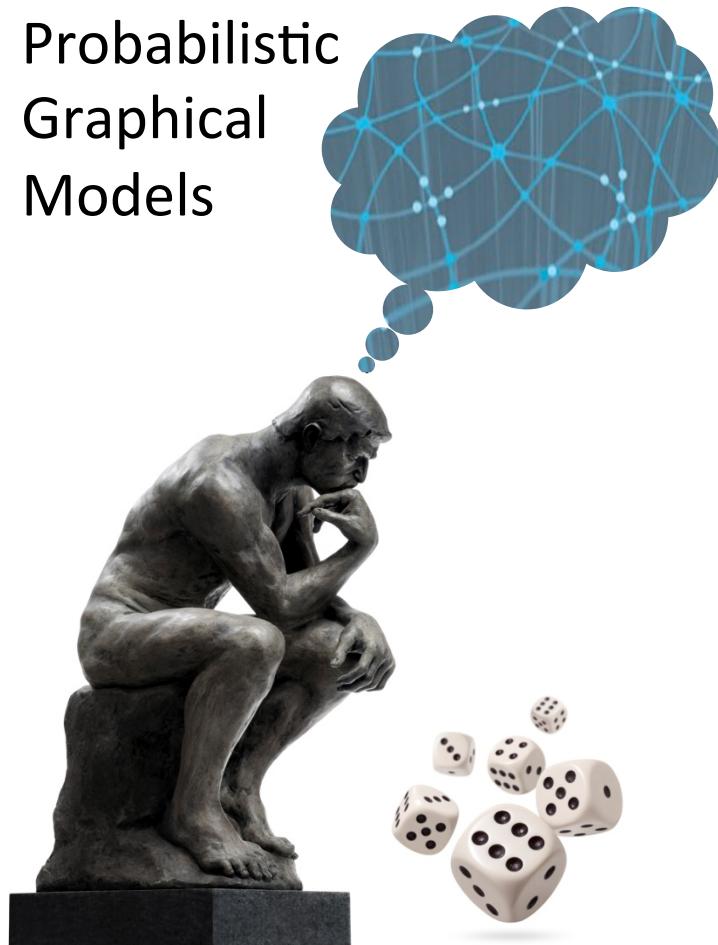


Probabilistic
Graphical
Models



Learning
Incomplete Data

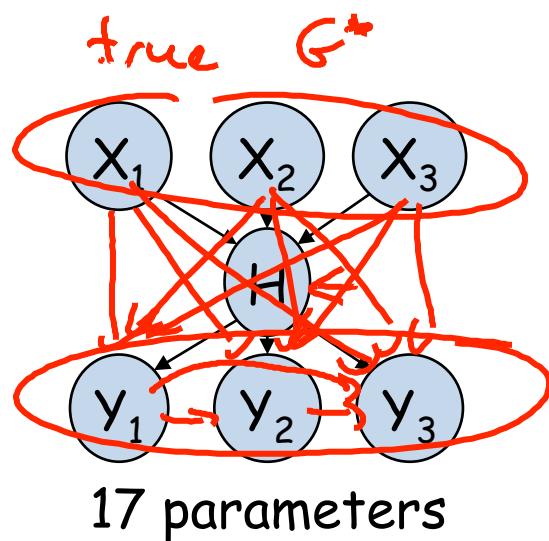
Overview

Incomplete Data

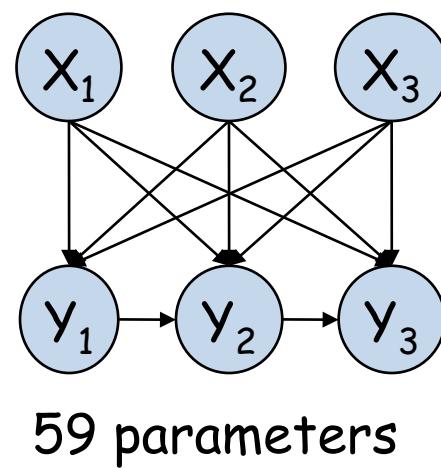
- Multiple settings:
 - Hidden variables
 - Missing values
- Challenges
 - Foundational – is the learning task well defined?
 - Computational – how can we learn with incomplete data?

Why latent variables?

- Model sparsity

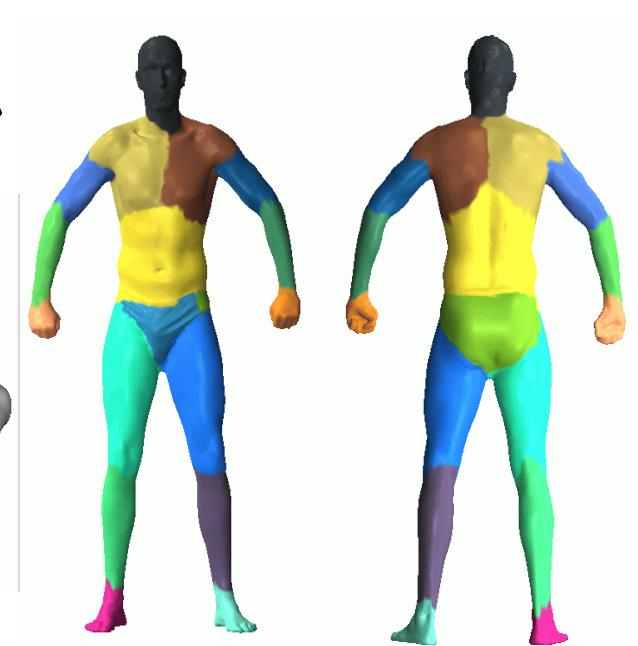
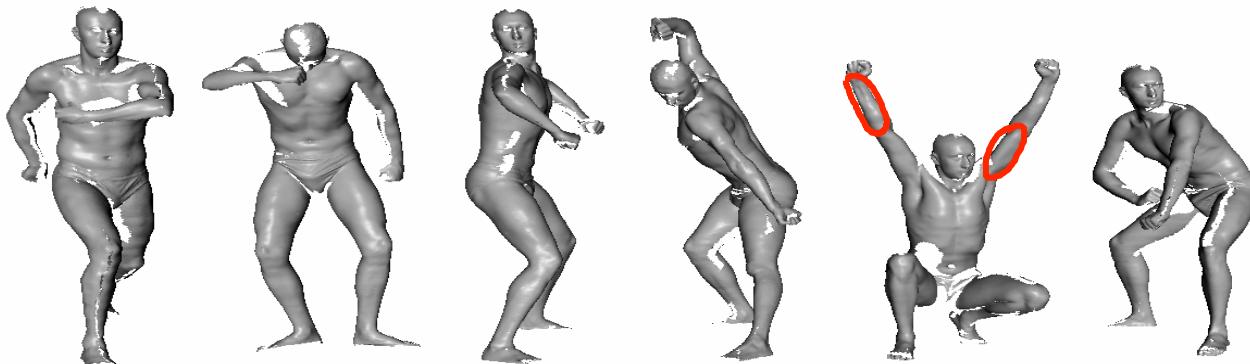


$p(y_1, y_2, y_3 | x_1, x_2, x_3)$



Why latent variables?

- Discovering clusters in data



Treating Missing Data

Sample sequence: H,T,?,?,H,?,H

- **Case I:** A coin is tossed on a table, occasionally it drops and measurements are not taken

H T ? ? H ? H

- **Case II:** A coin is tossed, but sometimes tails are not reported

H T T + H T H



We need to consider the missing data mechanism

Daphne Koller

Modeling Missing Data Mechanism

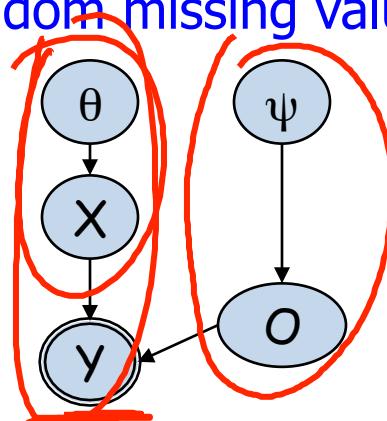
- $X = \{X_1, \dots, X_n\}$ are random variables
- $O = \{O_1, \dots, O_n\}$ are *observability variables*
 - Always observed $O_i = \begin{cases} 1 & X_i \text{ observed} \\ 0 & \text{otherwise} \end{cases}$
- $Y = \{Y_1, \dots, Y_n\}$ new random variables
 - $\text{Val}(Y_i) = \text{Val}(X_i) \cup \{?\}$
 - Always observed
 - Y_i is a deterministic function of X_i and O_i :

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

Modeling Missing Data Mechanism

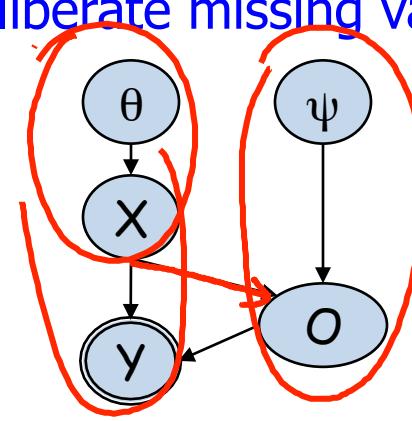
Case I

(random missing values)



Case II

(deliberate missing values)

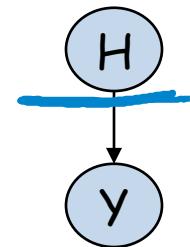


- When can we ignore the missing data mechanism and focus only on the likelihood?
- Missing at Random (MAR)

$$P_{\text{missing}} \models (O \perp H \mid d)$$

Identifiability

- Likelihood can have multiple global maxima
- Example:
 - We can rename the values of the hidden variable H
 - If H has two values, likelihood has two global maxima
- With many hidden variables, there can be an exponential number of global maxima
- Multiple local and global maxima can also occur with missing data (not only hidden variables)



Likelihood for Complete Data

Input Data:

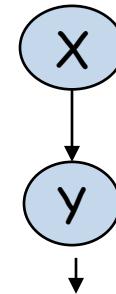
x	y
x^0	y^0
x^0	y^1
x^1	y^0

x^0	x^1
θ_{x0}	θ_{x1}

- Likelihood decomposes by variables
- Likelihood decomposes within CPDs

Likelihood:

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



x	$P(Y X)$	
	y^0	y^1
x^0	$\theta_{y^0 x^0}$	$\theta_{y^1 x^0}$
x^1	$\theta_{y^0 x^1}$	$\theta_{y^1 x^1}$

Likelihood for Incomplete Data

Input Data:

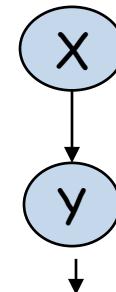
x	y
?	y^0
x^0	y^1
?	y^0

x^0	x^1
θ_{x0}	θ_{x1}

Likelihood:

- Likelihood does not decompose by variables
- Likelihood does not decompose within CPDs
- Computing likelihood requires inference!

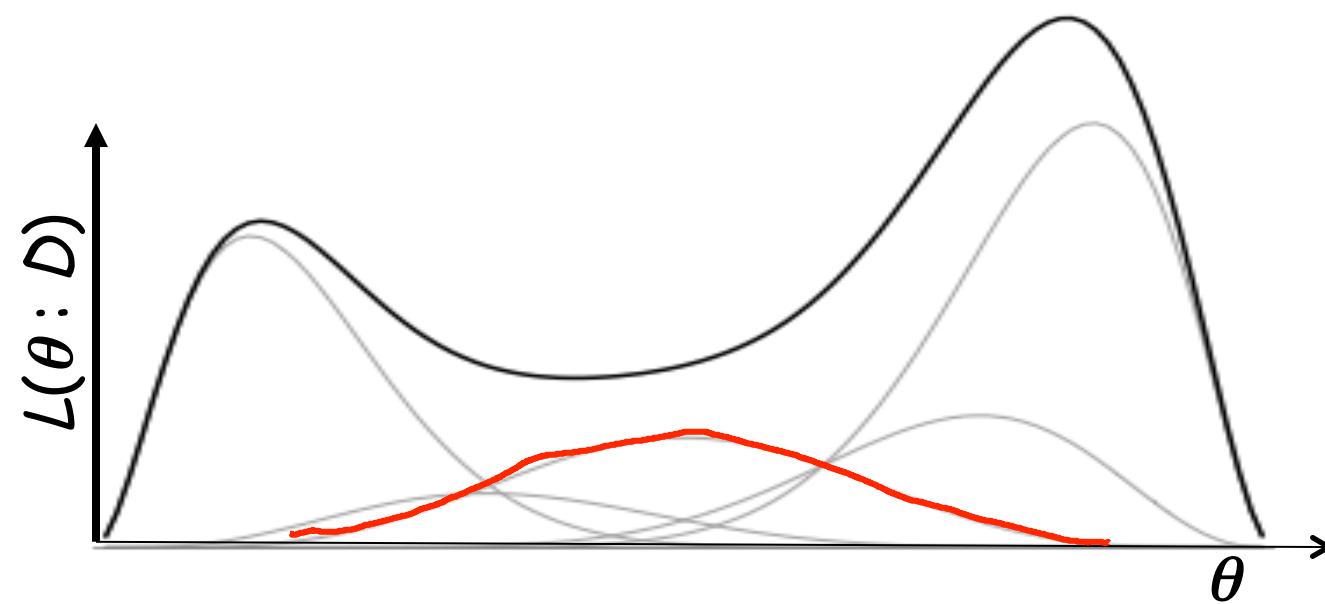
$$\begin{aligned}
 L(D : \theta) &= \underbrace{P(y^0)}_{\sum_{x \in Val(X)} P(x, y^0)} \cdot \underbrace{P(x^0, y^1)}_{\sum_{x \in Val(X)} P(x, y^1)} \cdot \underbrace{P(y^0)}_{\sum_{x \in Val(X)} P(x, y^0)} \\
 &= \left(\sum_{x \in Val(X)} P(x, y^0) \right)^2 \cdot P(x^0, y^1) \cdot \\
 &= \left(\theta_{x^0}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



x	$P(Y X)$	
	y^0	y^1
x^0	$\theta_{y^0 x^0}$	$\theta_{y^1 x^0}$
x^1	$\theta_{y^0 x^1}$	$\theta_{y^1 x^1}$

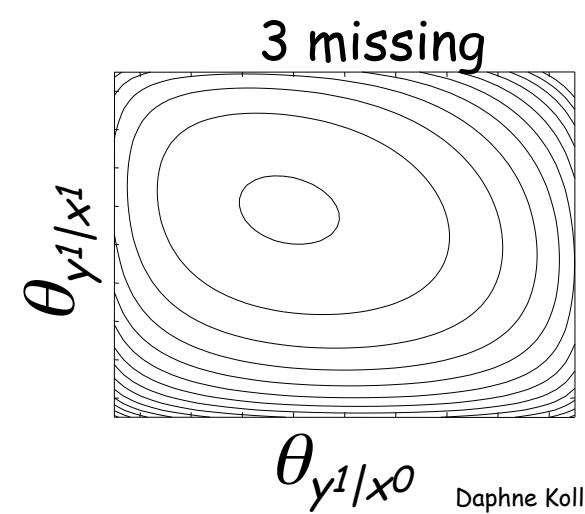
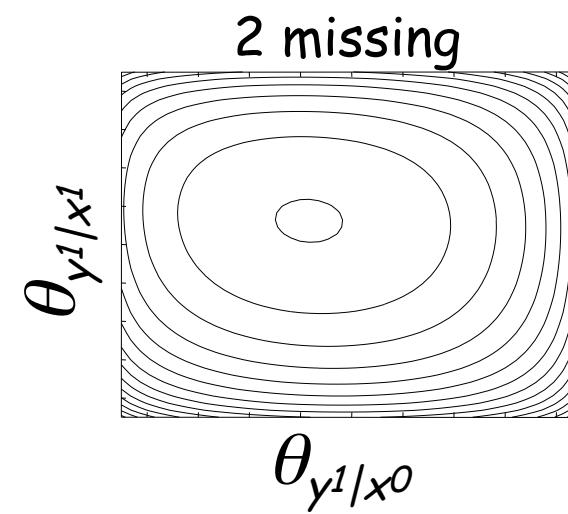
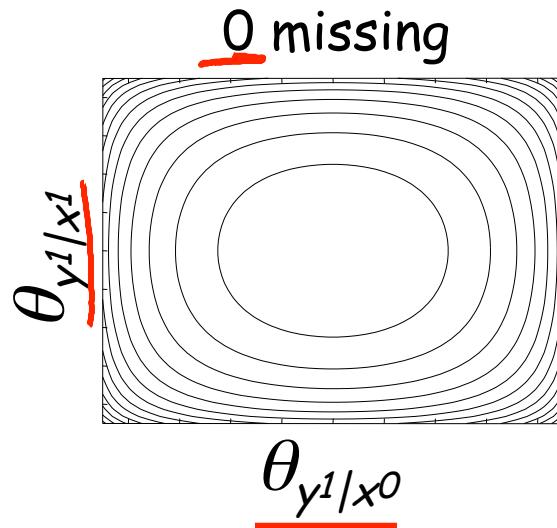
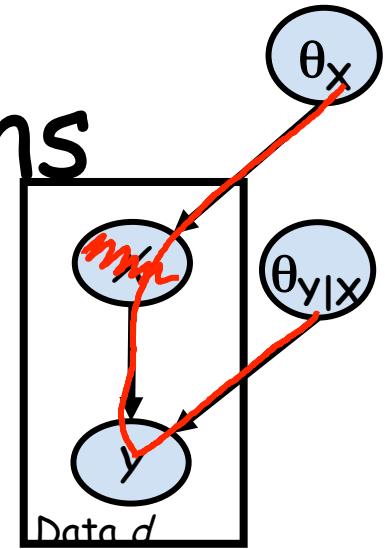
Daphne Koller

Multimodal Likelihood



Parameter Correlations

- Total of 8 data points
- Some X's unobserved

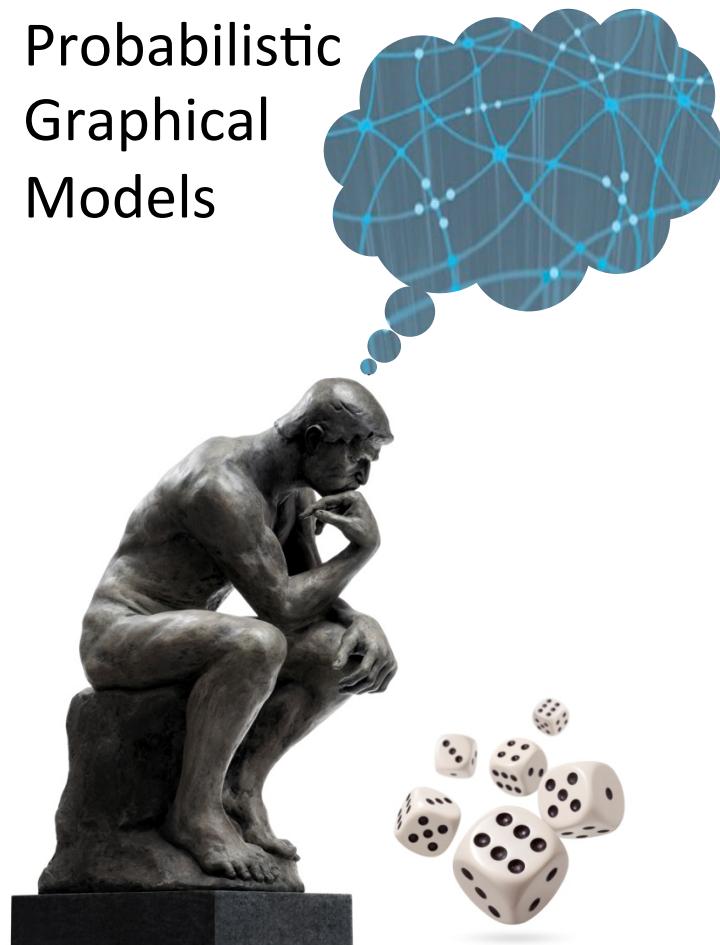


Daphne Koller

Summary

- Incomplete data arises often in practice
- Raises multiple challenges & issues:
 - The mechanism for missingness
 - Identifiability
 - Complexity of likelihood function

Probabilistic
Graphical
Models

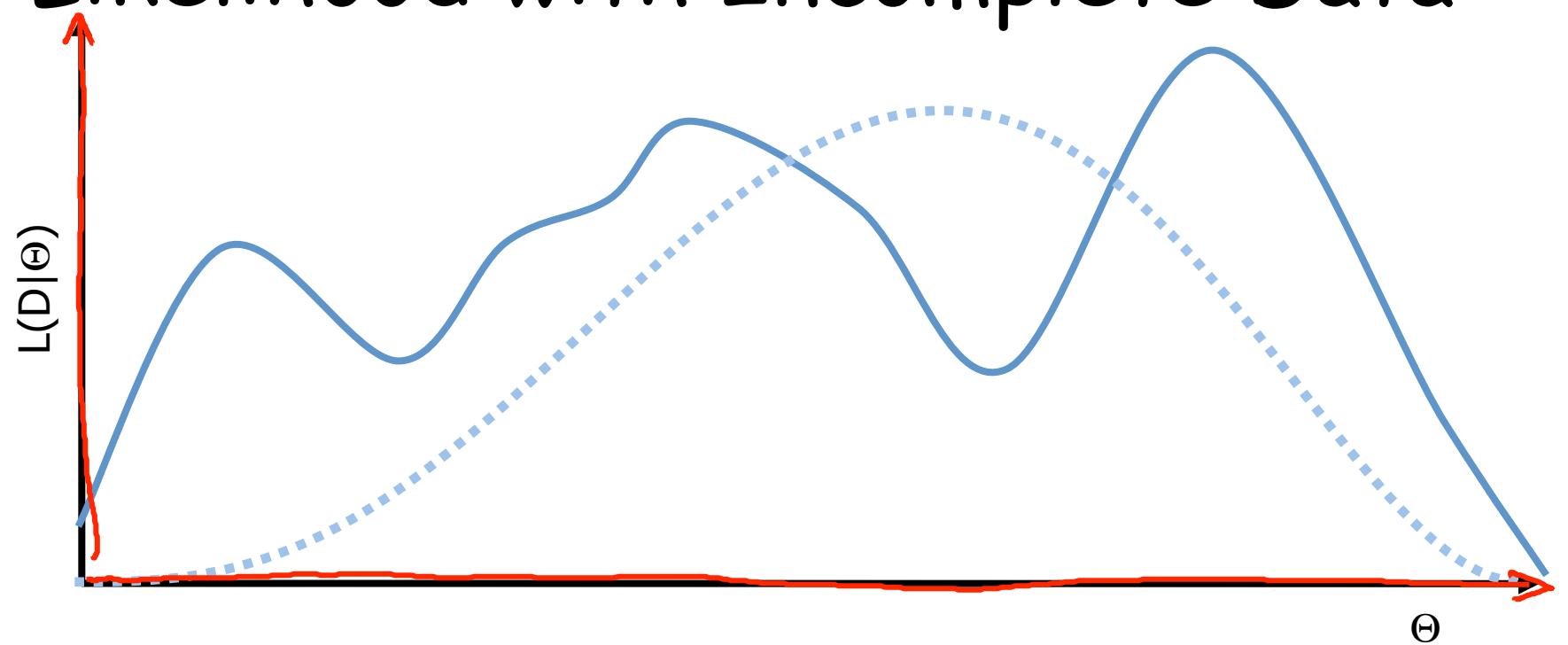


Learning

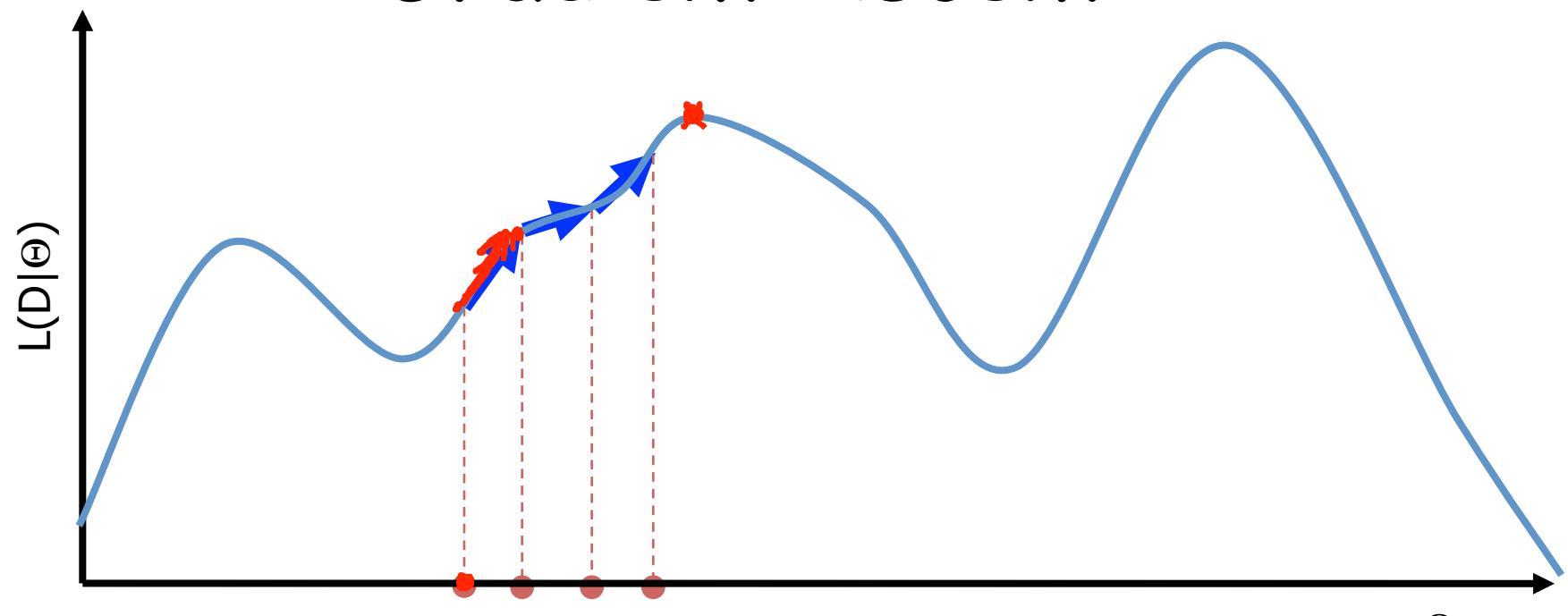
Incomplete Data

Likelihood
Optimization
Methods

Likelihood with Incomplete Data



Gradient Ascent



- Follow gradient of likelihood w.r.t. parameters
- Line search & conjugate gradient methods for fast convergence

Gradient Ascent

- Theorem:

$$\frac{\partial \log P(D | \Theta)}{\partial \theta_{x_i | u_i}} = \frac{1}{\theta_{x_i | u_i}} \sum_m P(x_i, u_i | d[m], \Theta)$$

evidence in mth instance
current param value
data instances m

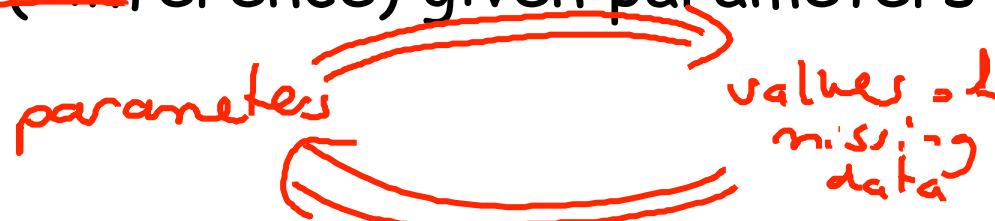
- Requires computing $P(X_i, U_i | d[m], \Theta)$ for all i, m
- Can be done with clique-tree algorithm, since X_i, U_i are in the same clique

Gradient Ascent Summary

- Need to run inference over each data instance at every iteration
 - Pros
 - Flexible, can be extended to non table CPDs
 - Cons
 - Constrained optimization: need to ensure that parameters define legal CPDs
 - For reasonable convergence, need to combine with advanced methods (conjugate gradient, line search)
- chain rule for derivatives*

Expectation Maximization (EM)

- Special-purpose algorithm designed for optimizing likelihood functions
- Intuition
 - Parameter estimation is easy given complete data
 - Computing probability of missing data is "easy" (=inference) given parameters



EM Overview

- Pick a starting point for parameters
- Iterate:
 - E-step (Expectation): “Complete” the data using current parameters
 - M-step (Maximization): Estimate parameters relative to data completion
(*Max Likelihood*)
- Guaranteed to improve $L(\theta : D)$ at each iteration

Expectation Maximization (EM)

- **Expectation (E-step):**

- For each data case $\underline{d[m]}$ and each family $\underline{X, U}$ compute
- Compute the expected sufficient statistics for each x, u

$$\overline{M}_{\theta^t}[x, \mathbf{u}] = \sum_{m=1}^M P(x, \mathbf{u} | \underline{d[m]}, \theta^t)$$

soft completion

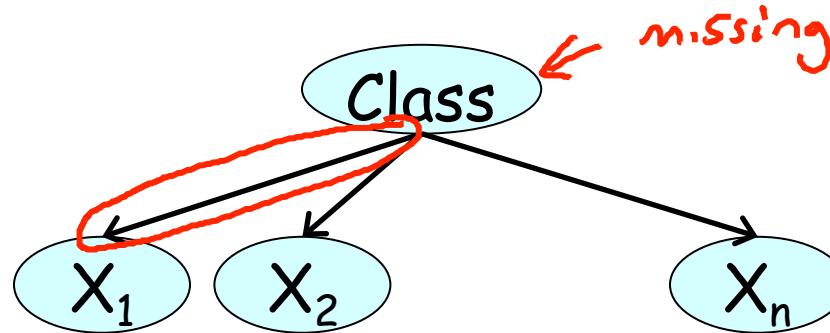
$$P(X, U | \underline{d[m]}, \theta^t)_{m|x, \bar{u}}$$

- **Maximization (M-step):**

- Treat the expected sufficient statistics (ESS) as if real
- Use MLE with respect to the ESS

$$\underline{\theta}_{x|\mathbf{u}}^{t+1} = \frac{\overline{M}_{\theta^t}[x, \mathbf{u}]}{\overline{M}_{\theta^t}[\mathbf{u}]}$$

Example: Bayesian Clustering



$$\begin{aligned}
 \bar{M}_{\theta}[c] &:= \sum_m P(c \mid \underbrace{x_1[m], \dots, x_n[m]}_{m}, \theta^t) & \theta_c^{t+1} &= \frac{\bar{M}_{\theta}[c]}{M} \\
 \bar{M}_{\theta}[x_i, c] &:= \sum_m P(c, x_i \mid \underbrace{x_1[m], \dots, x_n[m]}_{m}, \theta^t) & \theta_{x_i|c}^{t+1} &:= \frac{\bar{M}_{\theta}[x_i, c]}{\bar{M}_{\theta}[c]}
 \end{aligned}$$

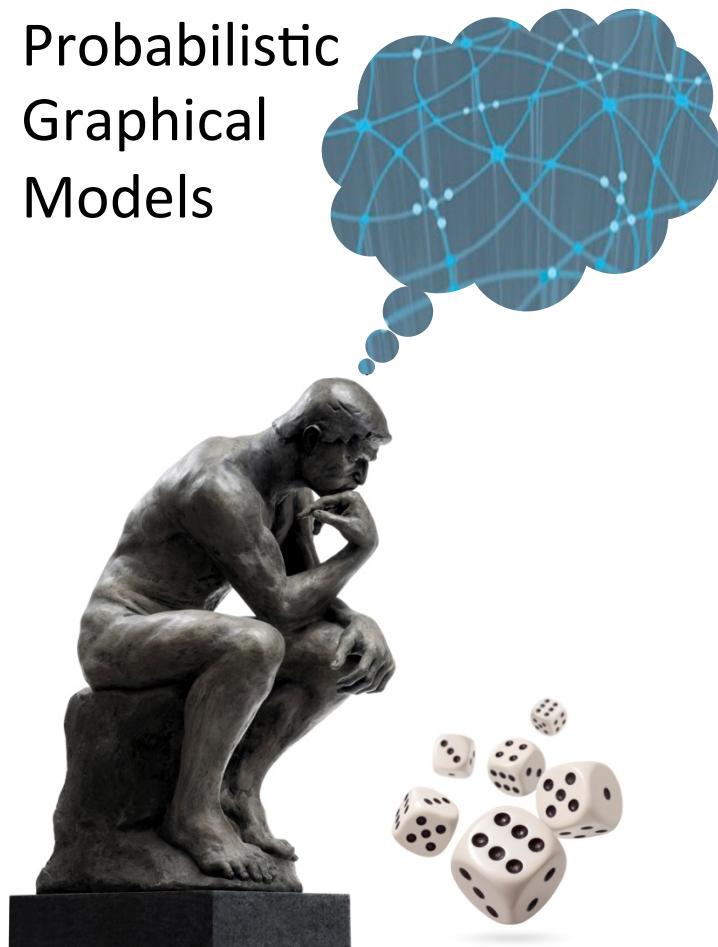
Annotations in red:

- A red arrow points from the word "missing" to the arrow from "Class" to X_n .
- Red underlines are placed under the summation variable m in both equations.
- Red arrows point from the θ terms in the equations to the corresponding terms in the definitions of \bar{M}_{θ} .
- A red arrow points from the term $\bar{M}_{\theta}[x_i, c]$ to the term $\theta_{x_i|c}^{t+1}$.

EM Summary

- Need to run inference over each data instance at every iteration
- Pros
 - Easy to implement on top of MLE for complete data
 - Makes rapid progress, especially in early iterations
- Cons
 - Convergence slows down at later iterations

Probabilistic
Graphical
Models

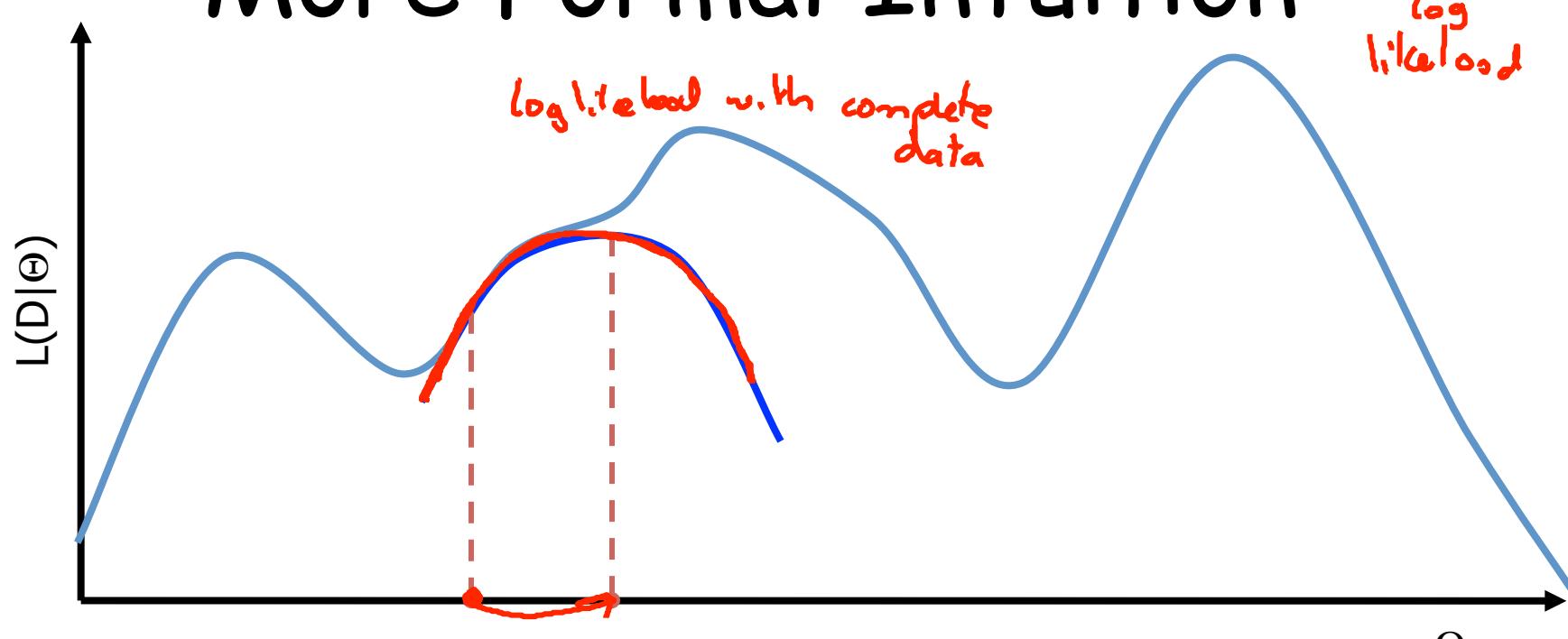


Learning

Incomplete Data

EM Analysis

More Formal Intuition



- Use current point to construct local approximation
- Maximize new function in closed form

More Formal Intuition

- d : observed data in instance
- H : hidden variables in instance
- $Q(H)$: distribution over hidden variables

$$\begin{aligned}
 \ell(\theta : \langle d, h \rangle) &= \sum_{i=1}^n \sum_{(x_i, u_i) \in Val(X_i, \text{Pa}_{X_i})} \underbrace{1_{\langle d, h \rangle}[x_i, u_i]}_{\text{assignment to } h} \log \theta_{x_i | u_i} \\
 E_{Q(H)}[\ell(\theta : \langle d, H \rangle)] &= \sum_{i=1}^n \sum_{(x_i, u_i) \in Val(X_i, \text{Pa}_{X_i})} \underbrace{E_{Q(H)}[1_{\langle d, H \rangle}[x_i, u_i]]}_{Q(x_i, u_i)} \log \theta_{x_i | u_i}
 \end{aligned}$$

↑ assignment to h
↑ log parameter

More Formal Intuition

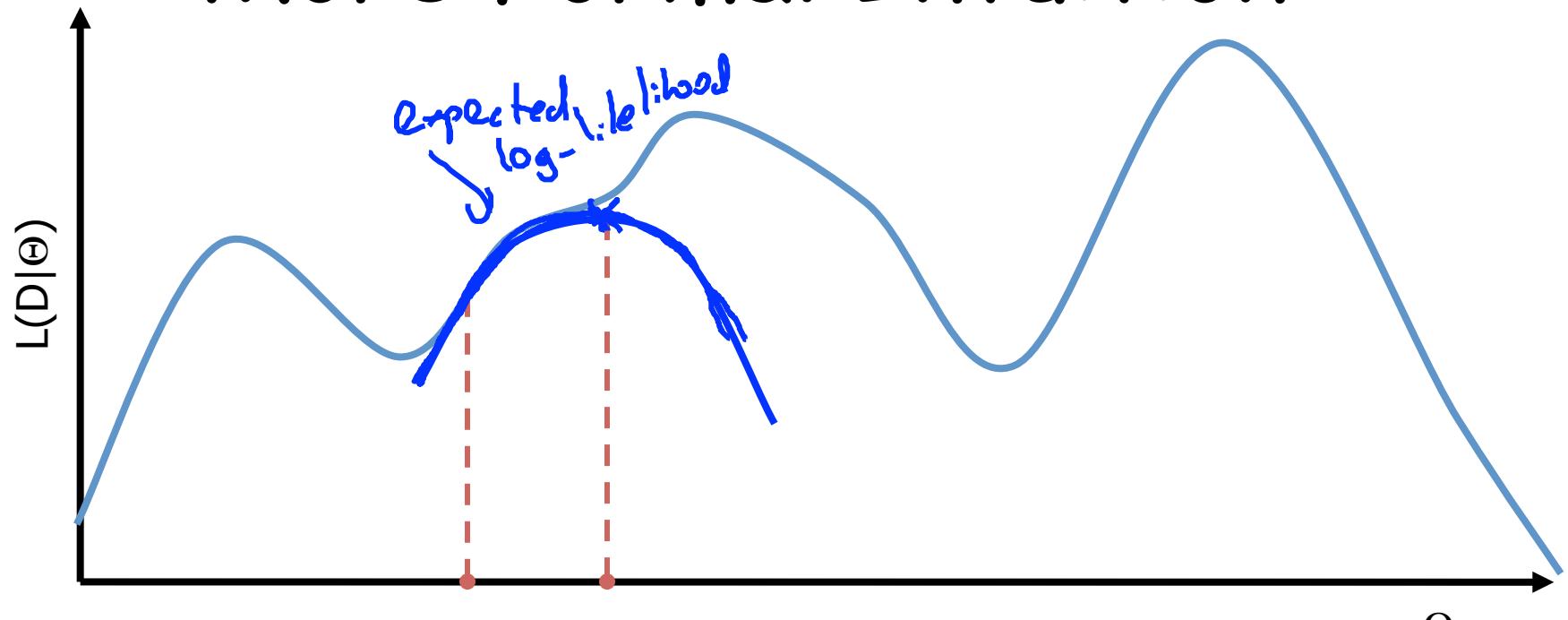
$$E_{Q(\mathbf{H})}[\ell(\boldsymbol{\theta} : \langle \mathbf{d}, \mathbf{H} \rangle)] = \sum_{i=1}^n \sum_{(x_i, \mathbf{u}_i)} Q(x_i, \mathbf{u}_i) \log \theta_{x_i | \mathbf{u}_i}$$

$$\underline{Q_m^t(\mathbf{H}[m])} = P(\mathbf{H}[m] \mid \mathbf{d}[m], \boldsymbol{\theta}^t)$$

$$\begin{aligned} & \sum_{m=1}^M \underline{E_{Q_m^t(\mathbf{H}[m])}[\ell(\boldsymbol{\theta} : \langle \mathbf{d}[m], \mathbf{H}[m] \rangle)]} \\ &= \sum_{i=1}^n \sum_{(x_i, \mathbf{u}_i)} \left(\sum_{m=1}^M \underline{P(x_i, \mathbf{u}_i \mid \mathbf{d}[m], \boldsymbol{\theta}^t)} \right) \log \theta_{x_i | \mathbf{u}_i} \\ &= \sum_{i=1}^n \sum_{(x_i, \mathbf{u}_i)} \underline{\bar{M}_{\boldsymbol{\theta}^t}[x_i, \mathbf{u}_i]} \log \theta_{x_i | \mathbf{u}_i} \end{aligned}$$

expected suff stats
 log likelihood for
 complete data
 using ESS

More Formal Intuition



- Use current point to construct local approximation
- Maximize new function in closed form

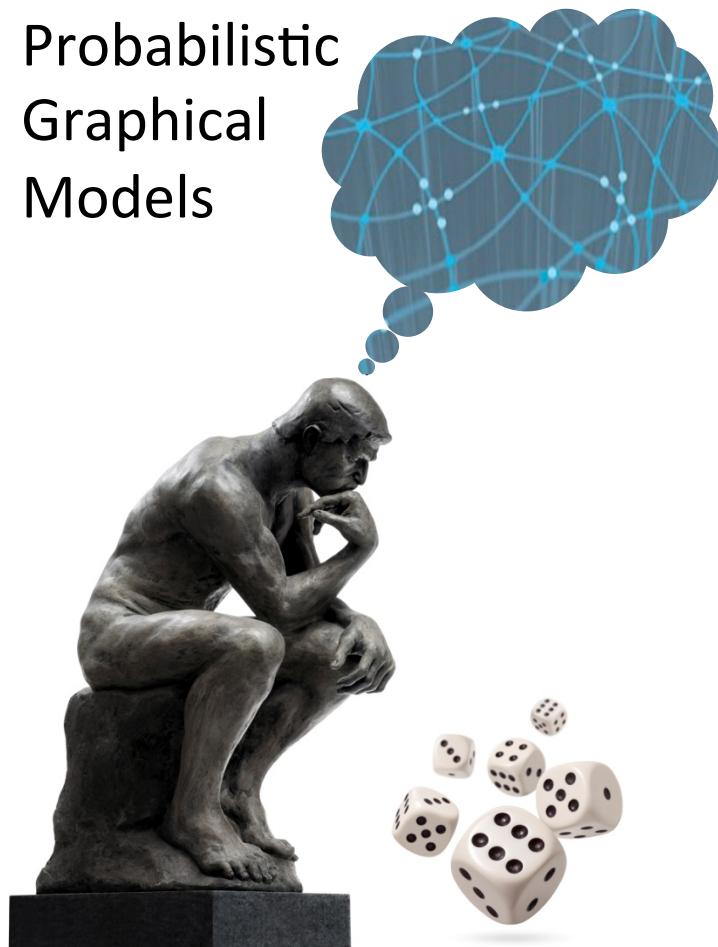
EM Guarantees

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

gradient is zero



Probabilistic
Graphical
Models

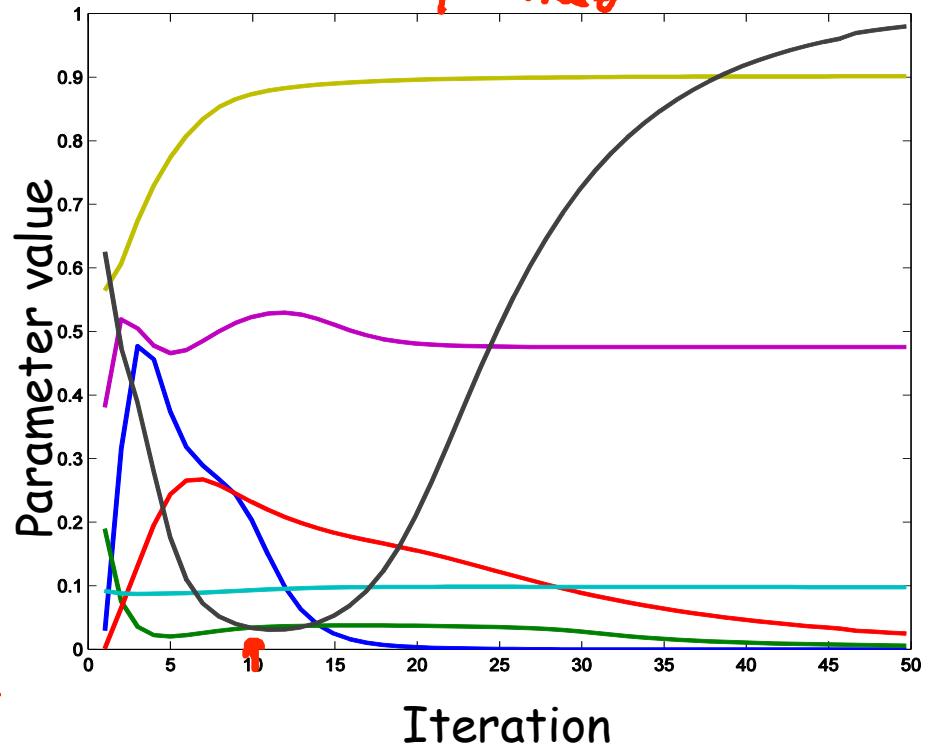
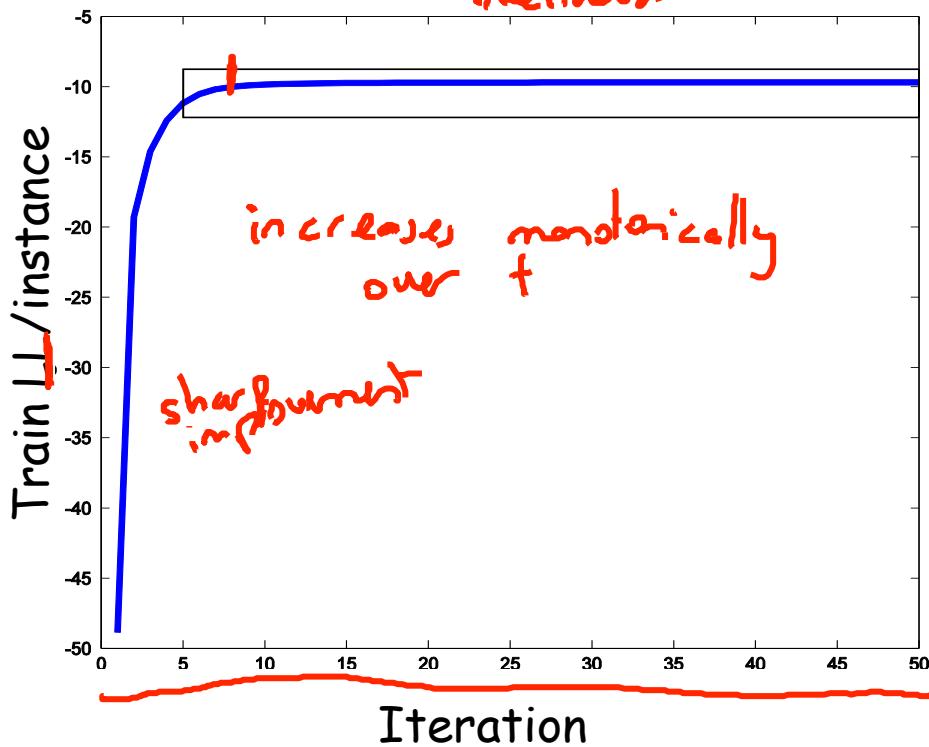


Learning

Incomplete Data

EM in Practice

EM Convergence in Practice

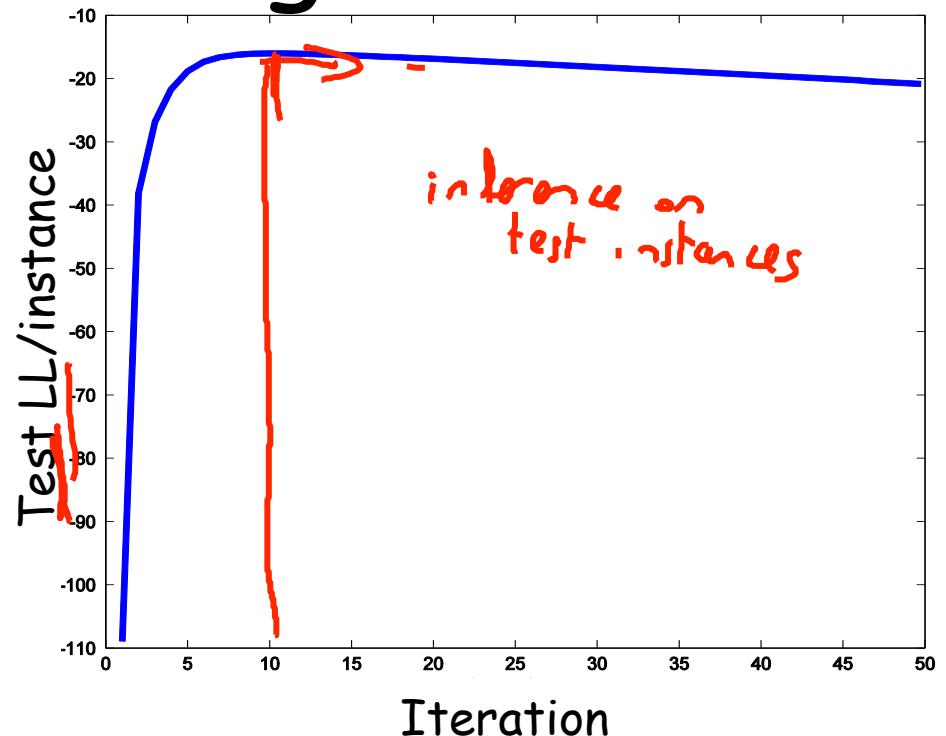
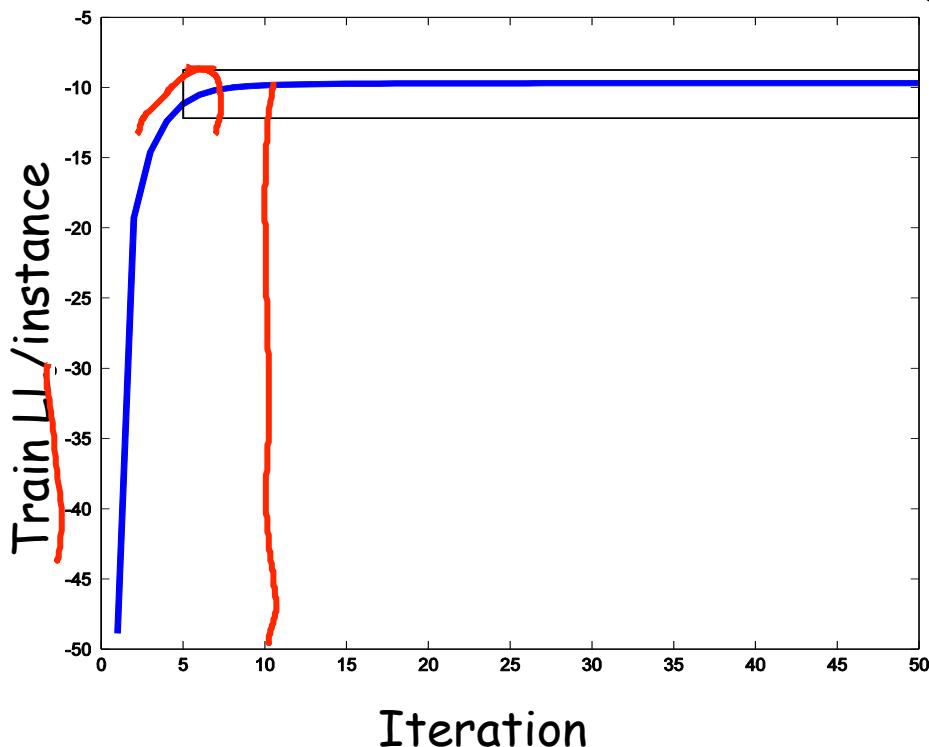


G. Elidan

Daphne Koller

Overfitting

(numerical)

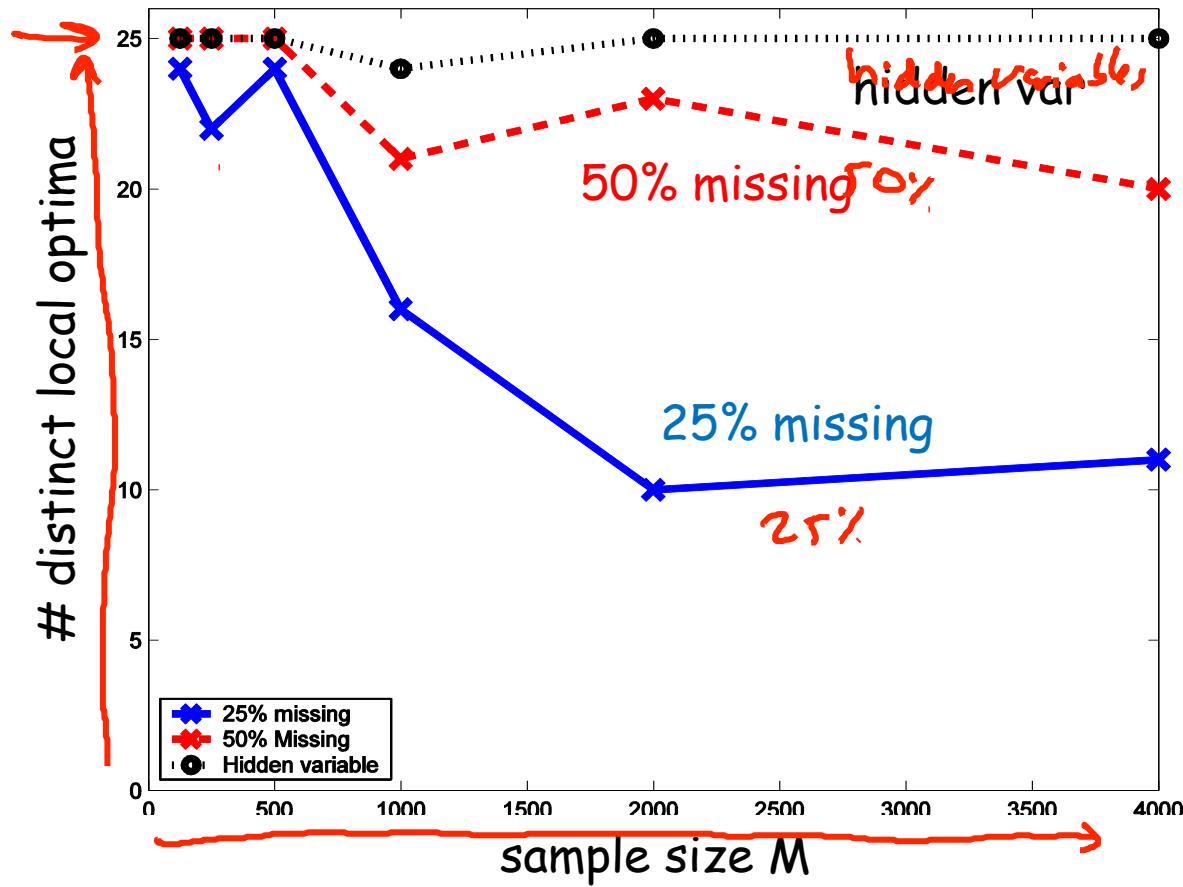


- Early stopping using cross validation
- Use MAP with parameter priors rather than MLE

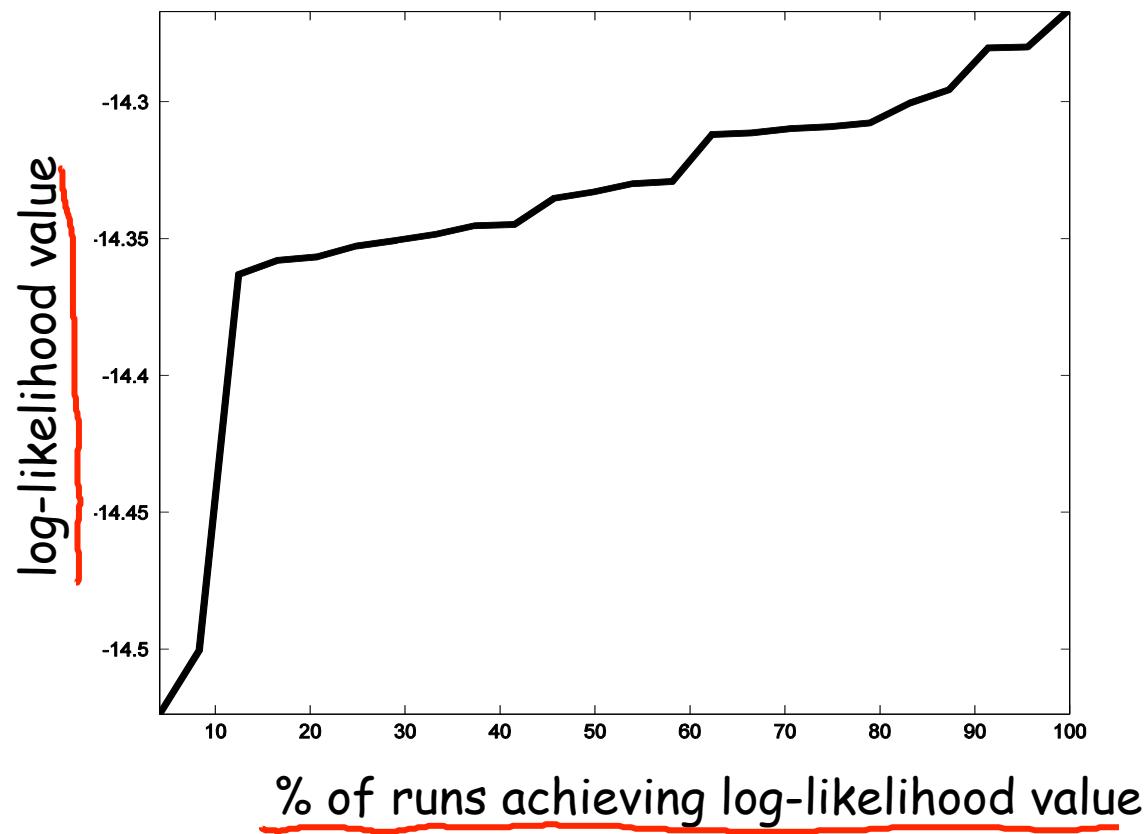
G. Elidan

Daphne Koller

Local Optima



Significance of Local Optima



G. Elidan

% of runs achieving log-likelihood value

Daphne Koller

Initialization is Critical

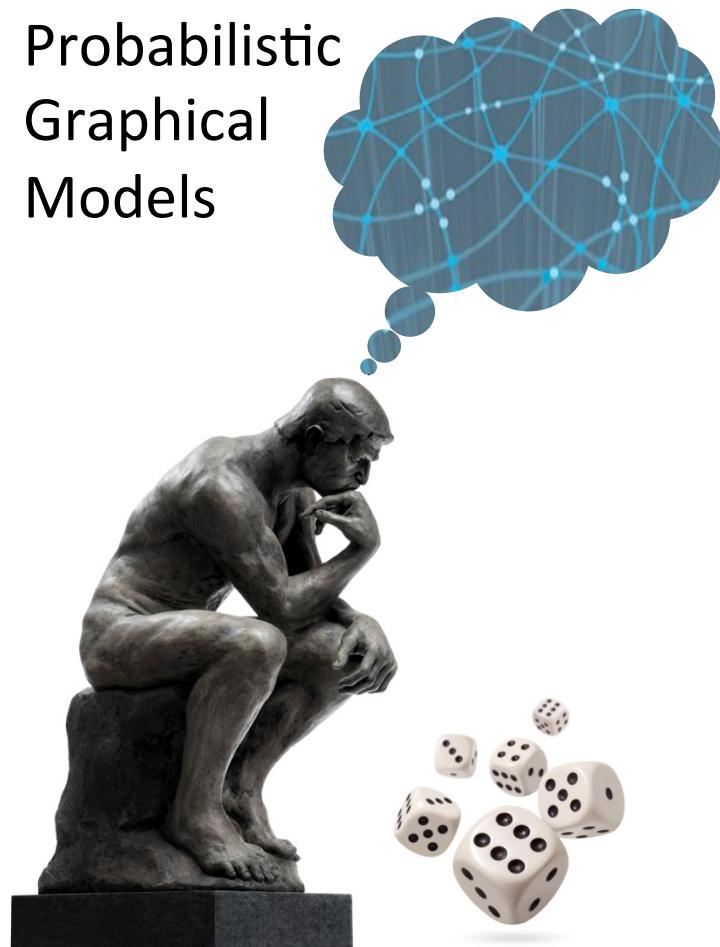
- Multiple random restarts
- From prior knowledge
- From the output of a simpler algorithm

clustering (k-means
hierarchical
agglomerative
clustering)

Summary

- Convergence of likelihood \neq convergence of parameters
- Running to convergence can lead to overfitting
- Local optima are unavoidable, and increase with the amount of missing data
- Local optima can be very different
- Initialization is critical

Probabilistic
Graphical
Models

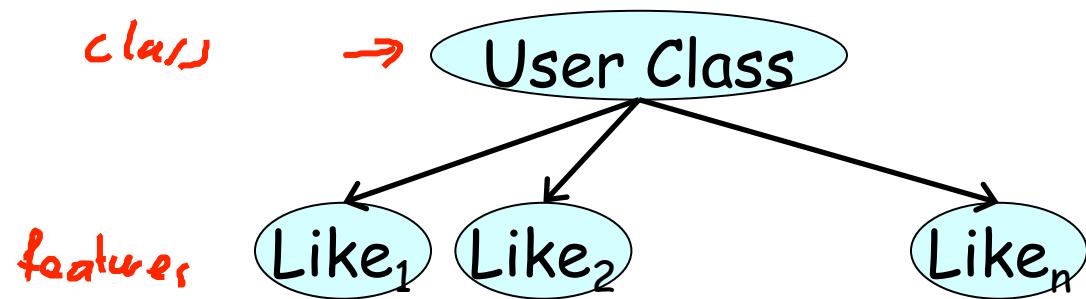


Learning

Incomplete Data

Learning with
Latent
Variables

Discovering User Clusters



J. Breese

MSNBC Story clusters

(Readers of commerce and technology stories (36%):

- E-mail delivery isn't exactly guaranteed
- Should you buy a DVD player?
- Price low, demand high for Nintendo

Sports Readers (19%):

- Umps refusing to work is the right thing
- Cowboys are reborn in win over eagles
- Did Orioles spend money wisely?

Readers of top promoted stories (29%):

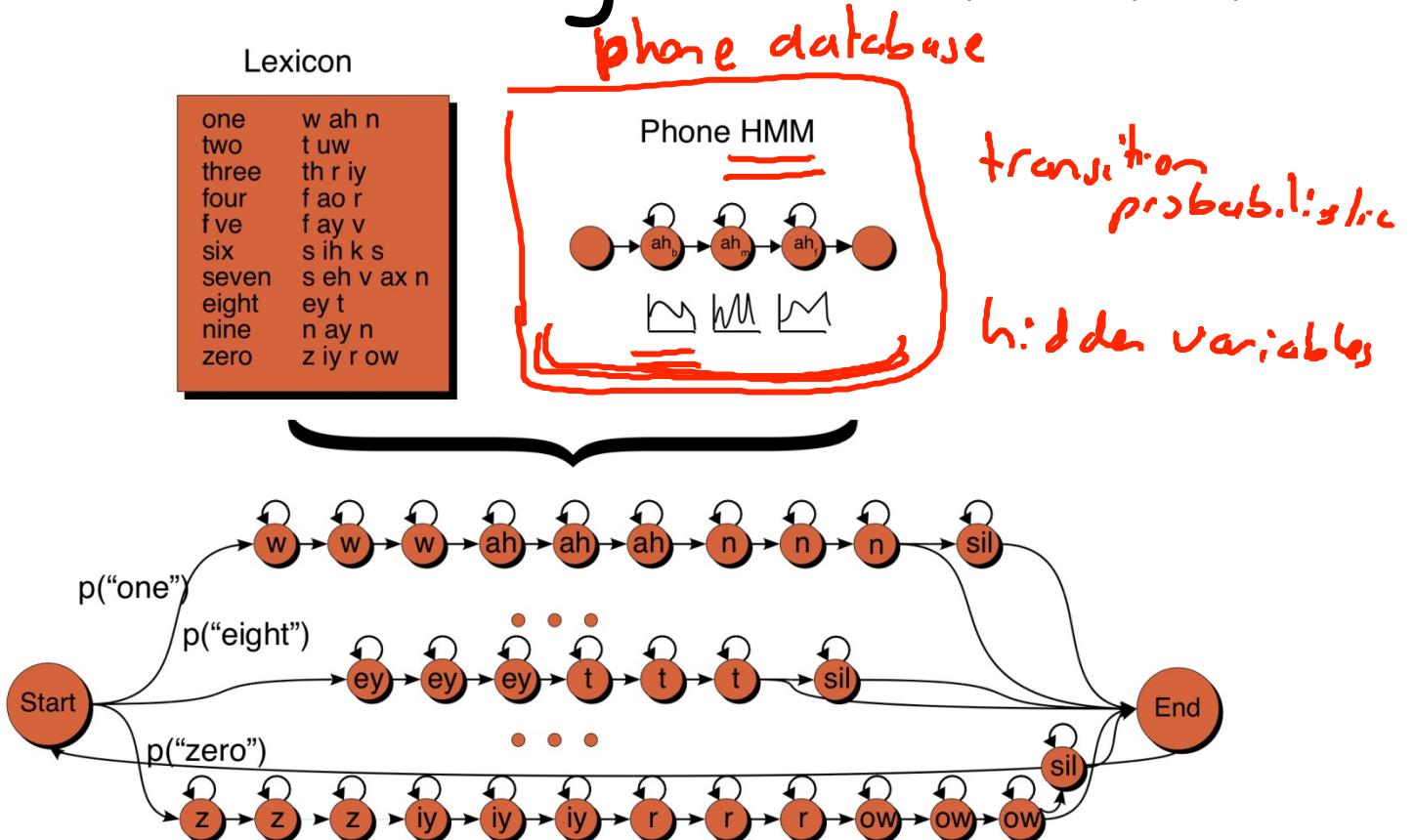
- 757 Crashes At Sea
- (Israel, Palestinians Agree To Direct Talks
- (Fuhrman Pleads Innocent To Perjury

Readers of "Softer" News (12%):

- The truth about what things cost
- Fuhrman Pleads Innocent To Perjury
- Real Astrology

Daphne Koller

Speech Recognition HMM



Dan Jurafsky, Stanford

Daphne Koller

3D Robot Mapping

- Input: Point cloud from laser range finder obtained by moving robot
 - Output: 3D planar map of environment
- • Parameters: Location & angle of walls (planes)
- • Latent variables: Assignment of points to walls
association

Thrun, Martin, Liu, Haehnel, Emery-Montemerlo, Chakrabarti, Burgard,
IEEE Transactions on Robotics, 2004

Daphne Koller

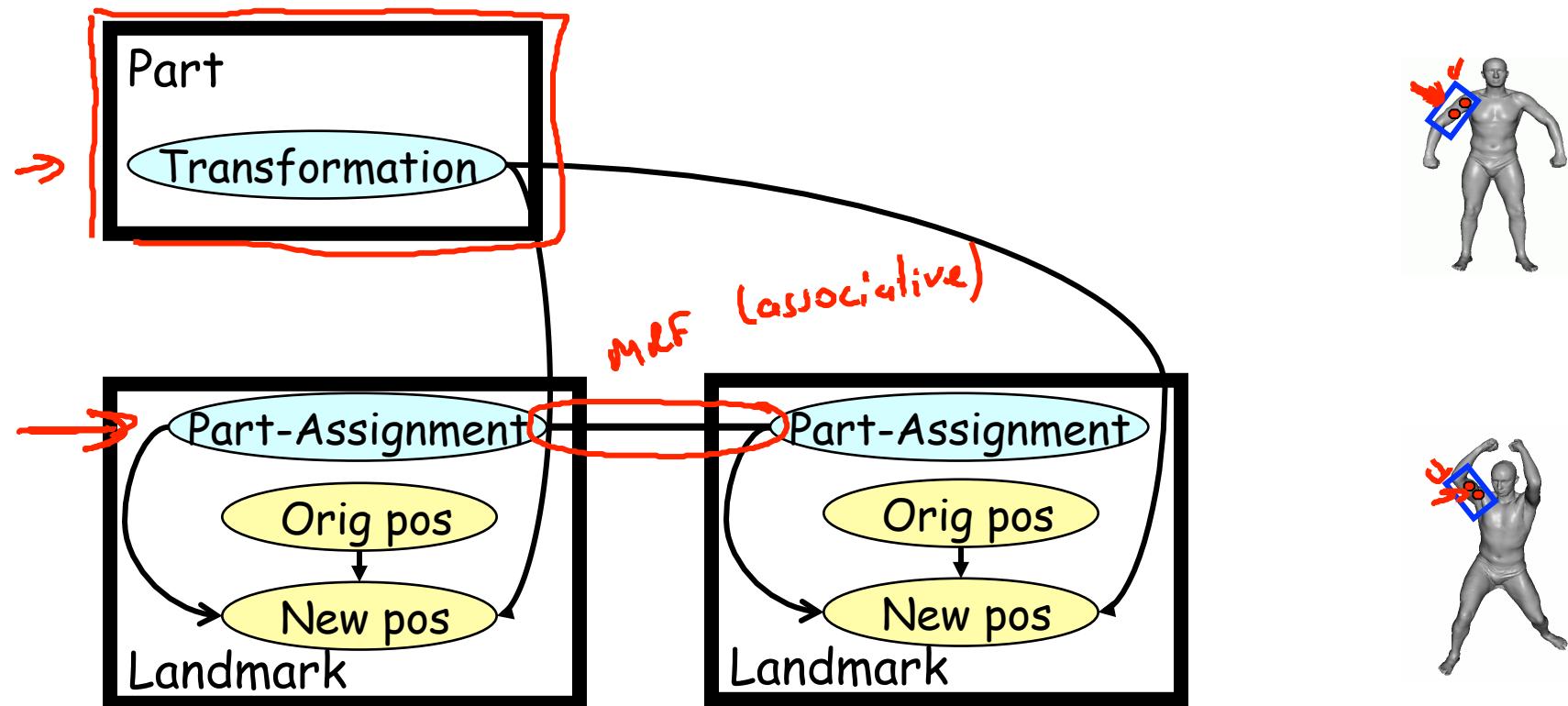
Body Parts from Point Cloud Scans



Anquellov, Koller, Pang, Srinivasan, Thrun UAI 2004

Daphne Koller

Collective Clustering Model



Anquellov, Koller, Pang, Srinivasan, Thrun UAI 2004

Daphne Koller



Anquellov, Koller, Pang, Srinivasan, Thrun UAI 2004

Daphne Koller

Helicopter Demo Alignment

- Input: Multiple sample trajectories by different pilots flying same sequence
- Output:
 - Aligned trajectories
 - Model of "template" trajectory



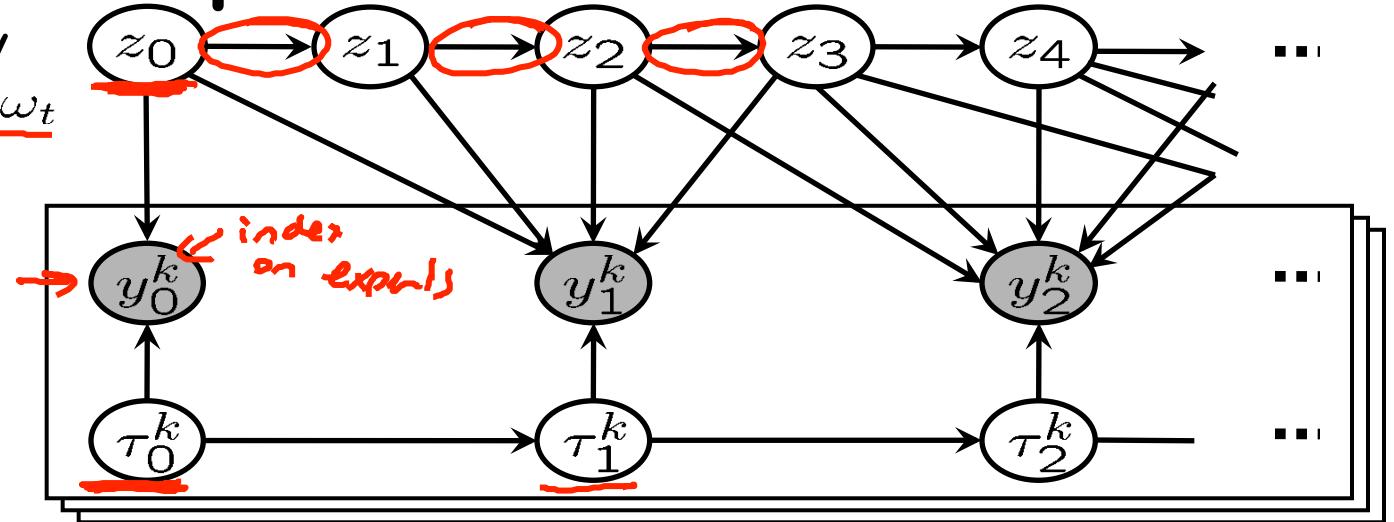
Coates, Abbeel, Ng, ICML 2008

Daphne Koller

Graphical model

Intended trajectory
 $z_{t+1} = \underline{f(z_t) + \omega_t}$

Expert demonstrations
 $y_j = z_{\tau_j} + \nu_j$
Time indices



Picking Latent Variable Cardinality

- If we use likelihood for evaluation, more values is always better
- Can use score that penalizes complexity
 - BIC - tends to underfit
 - Extensions of BDe to incomplete data (approximations)
- Can use metrics of cluster coherence to decide whether to add/ remove clusters
- Bayesian methods (Dirichlet processes) can average over different cardinalities
(MCMC) distribution over cardinality

Summary

- Latent variables are perhaps the most common scenario for incomplete data
 - often a critical component in constructing models for richly structured domains
- Latent variables satisfy MAR, so can use EM
- Serious issues with unidentifiability & multiple optima
necessitate good initialization
- Picking variable cardinality is a key question