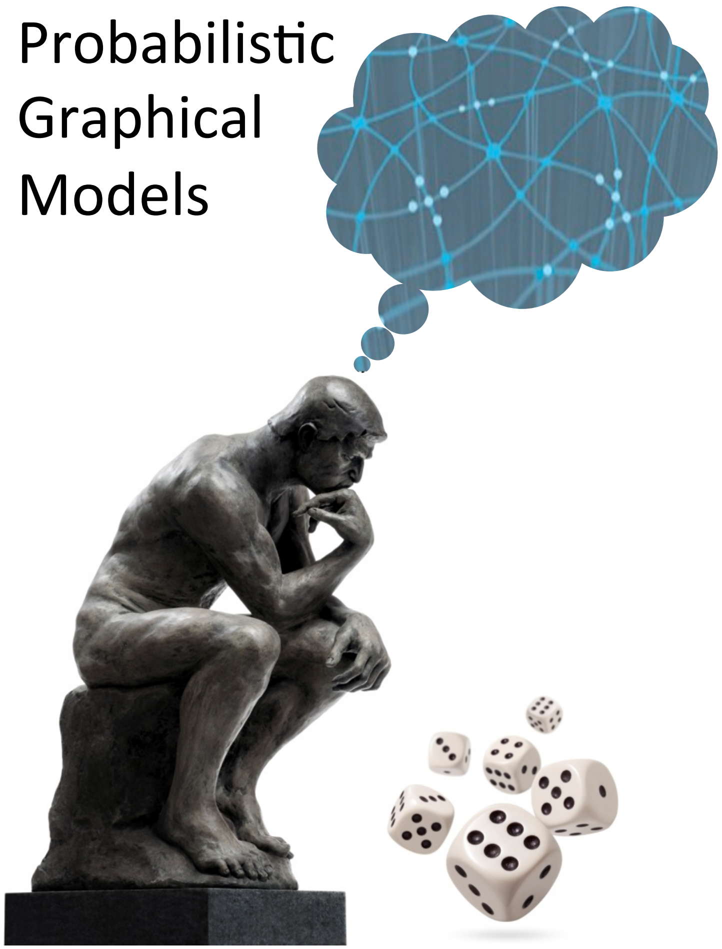
Probabilistic Graphical Models



Learning

Incomplete Data

Overview

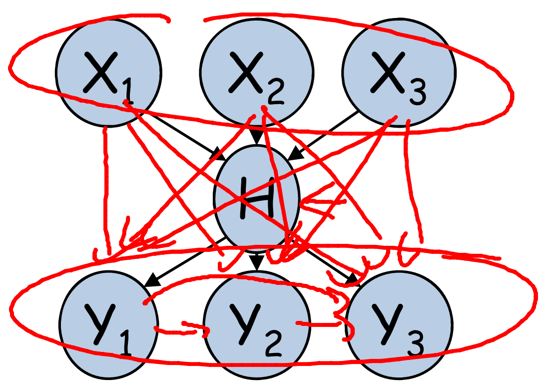
Incomplete Data

- 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?

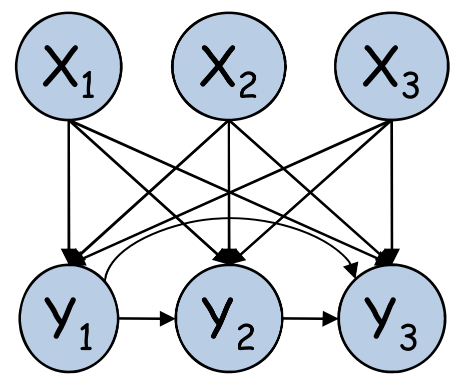
- Model sparsity

true G^*



17 parameters

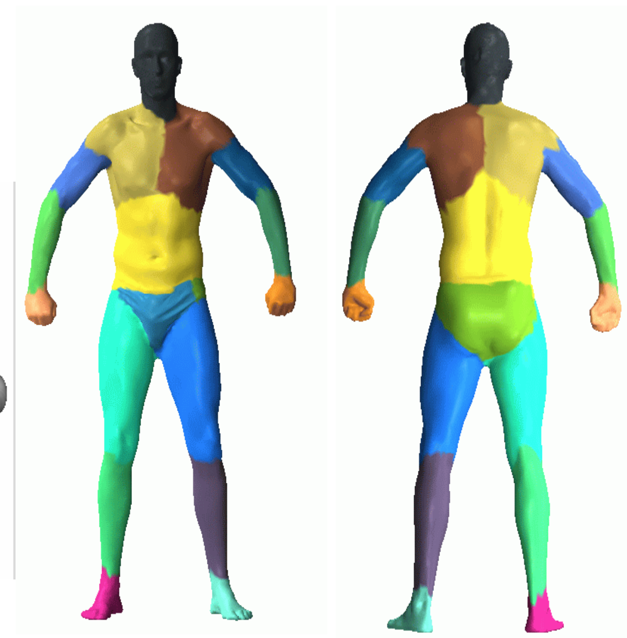
$p(x_1, x_2, x_3, y_1, y_2, y_3)$



59 parameters

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 T H T H

We need to consider the missing data mechanism

Modeling Missing Data Mechanism

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

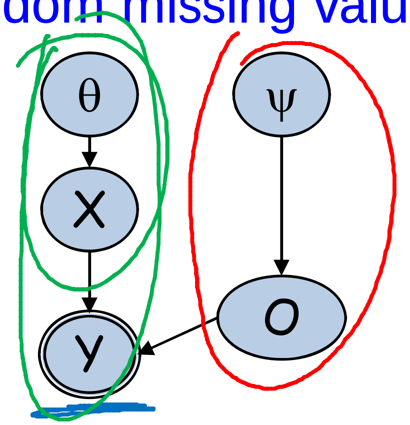
$$Y_i = \begin{cases} X_i & O_i = o^1 \\ ? & O_i = o^0 \end{cases}$$

Daphne Koller

Modeling Missing Data Mechanism

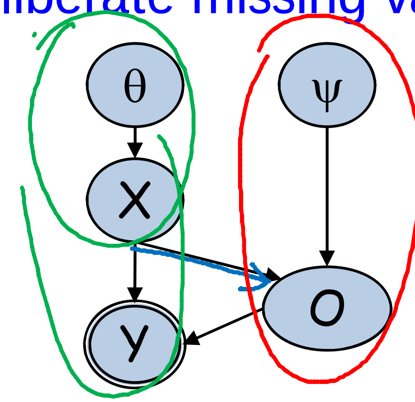
Case I

(random missing values)



Case II

(deliberate missing values)



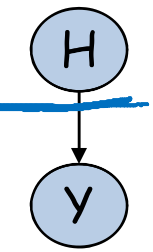
- When can we ignore the missing data mechanism and focus only on the likelihood?
- Missing at Random (MAR)

unobserved X's

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

observed values X

Identifiability



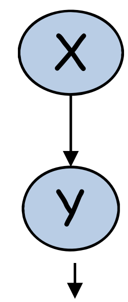
- Likelihood can have multiple global maxima
- Example:
  - We can rename the values of the hidden variable H
  - If H has two values, likelihood has two global maxima
- With many hidden variables, there can be an exponential number of global maxima
- Multiple local and global maxima can also occur with 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}



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 decomposes by variables
- Likelihood decomposes within CPDs

Likelihood:

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

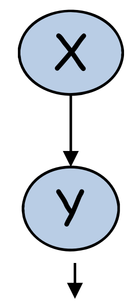
Daphne Koller

Likelihood for Incomplete Data

Input Data:

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

x^0	x^1
θ_{x^0}	θ_{x^1}



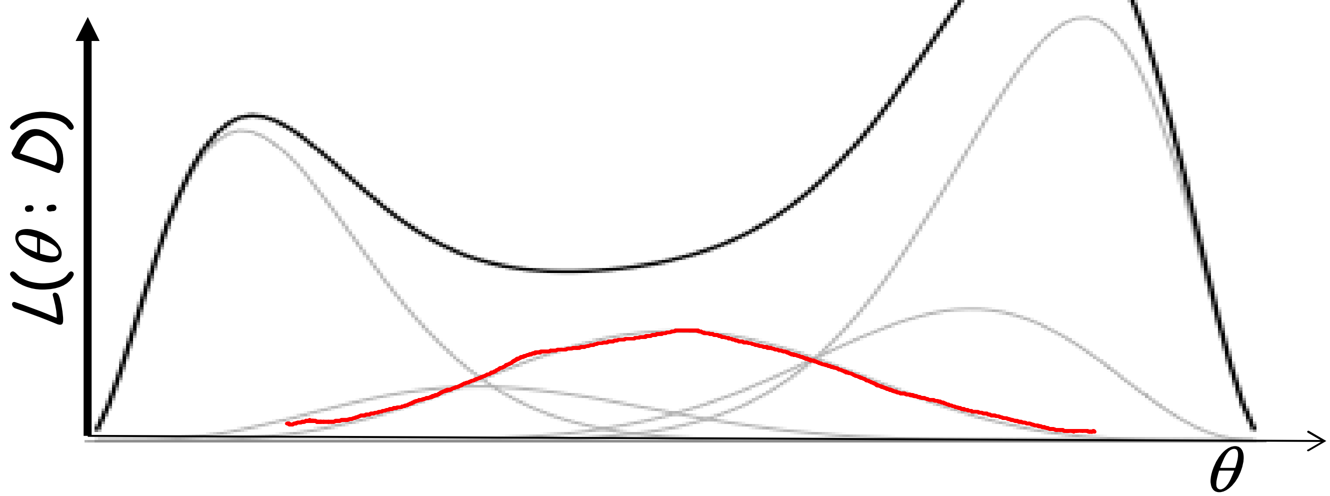
X	*P(Y/X)* y^0	y^1
x^0	$\theta_{y^0\mid x^0}$	$\theta_{y^1\mid x^0}$
x^1	$\theta_{y^0\mid x^1}$	$\theta_{y^1\mid x^1}$

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

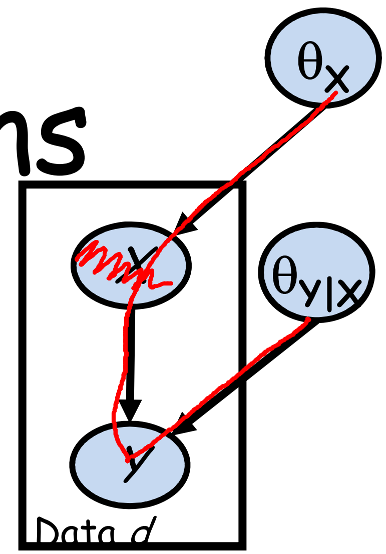
Likelihood:

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

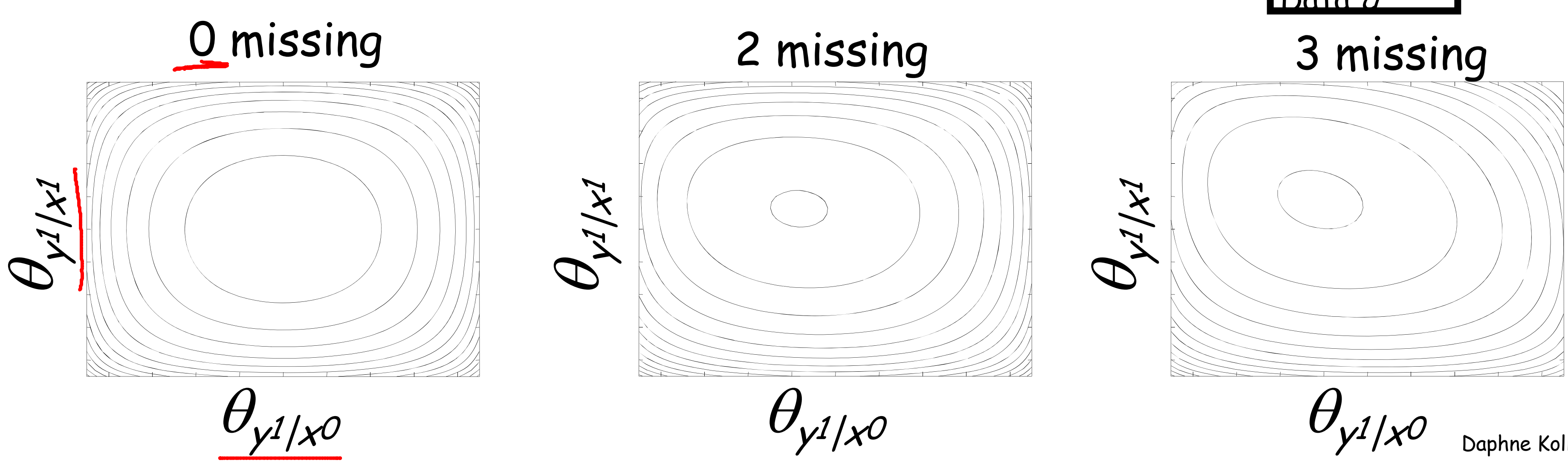
Multimodal Likelihood



Parameter Correlations



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

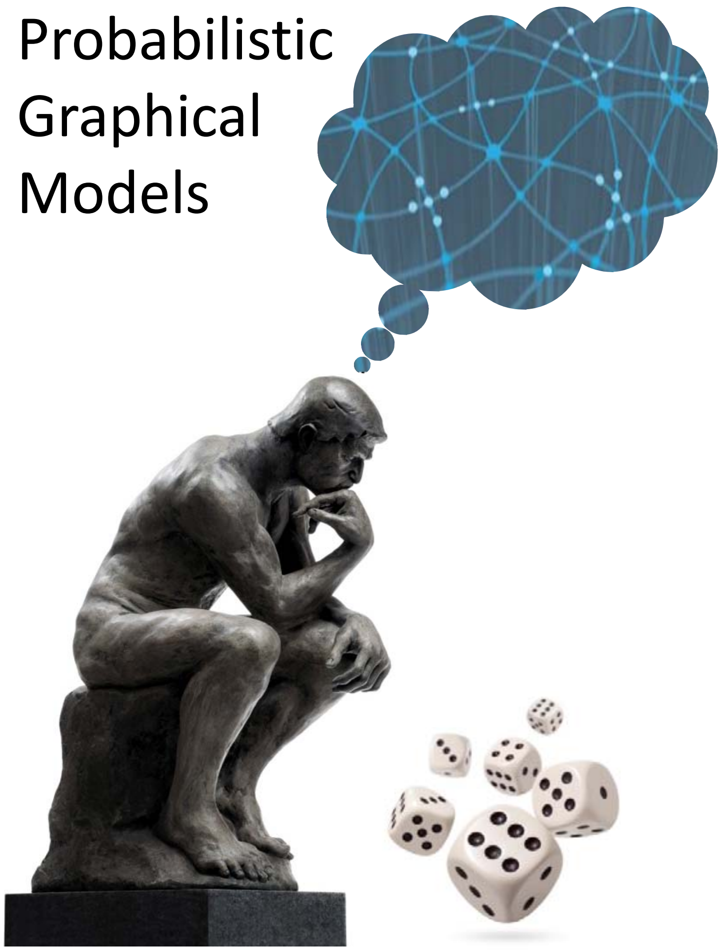


Daphne Koller

Summary

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

Probabilistic
Graphical
Models

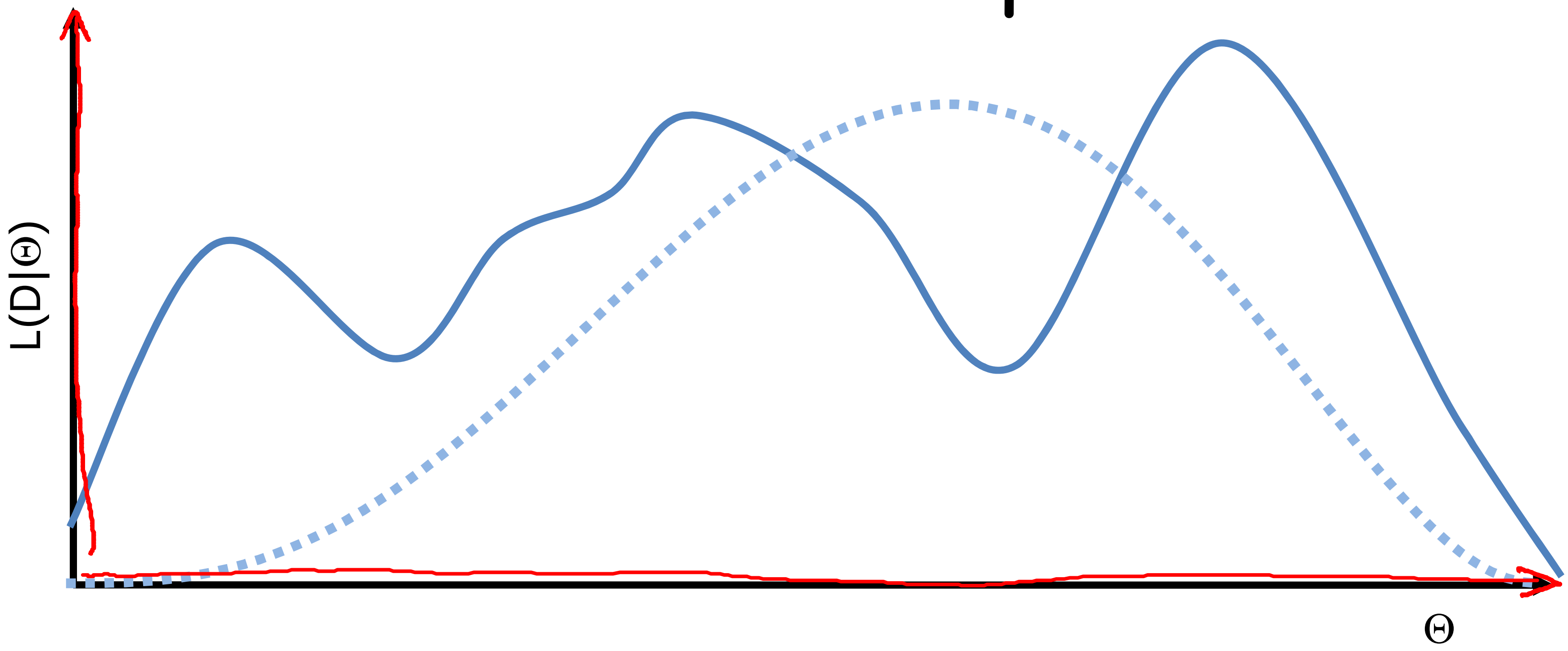


Learning

Incomplete Data

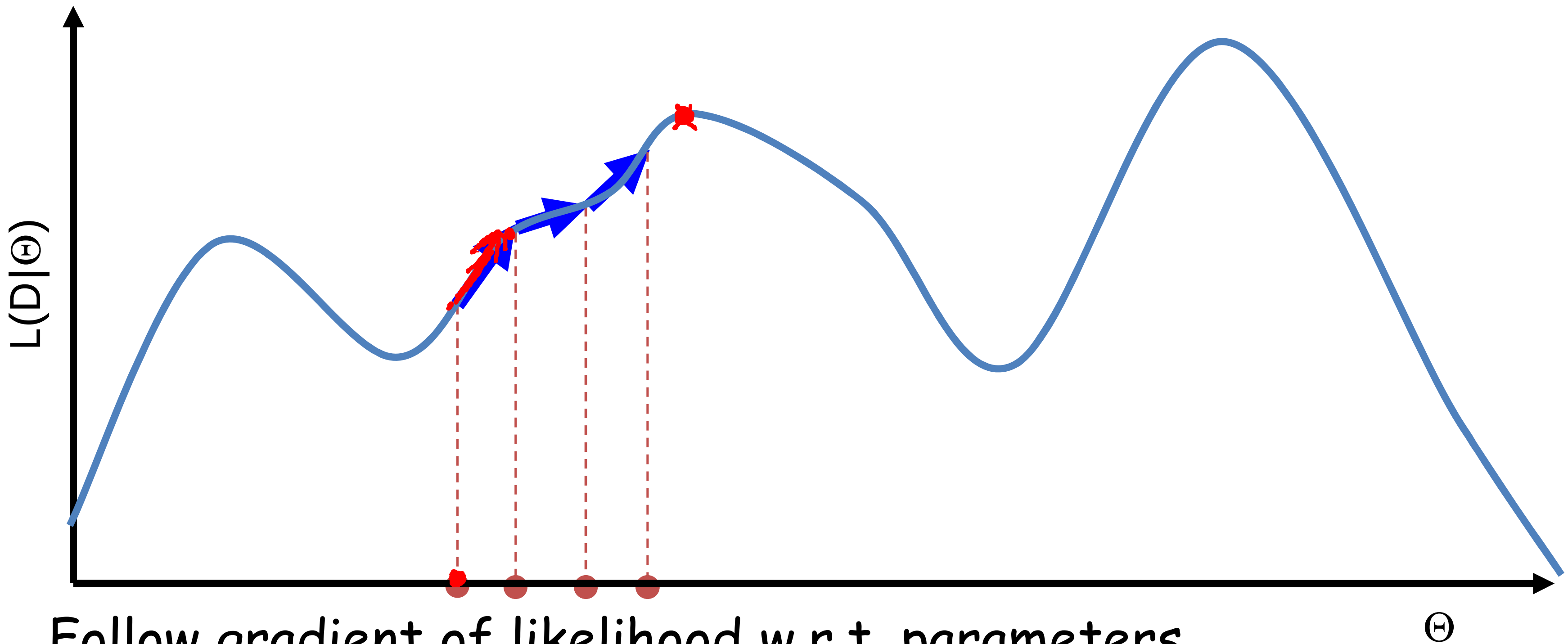
Likelihood Optimization Methods

Likelihood with Incomplete Data



Daphne Koller

Gradient Ascent



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

Gradient Ascent

- Theorem:

evidence in m^{th} instance

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

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

    chain rule for derivatives
- Cons
  - Constrained optimization: need to ensure that parameters define legal CPDs
  - For reasonable convergence, need to combine with advanced methods (conjugate gradient, line search)

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

parameters

values of missing data

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 **d**[m] and each family X,**U** compute

    soft completion

    $$P(X,\boldsymbol{U} \mid \mathbf{d}[m],\theta^t)$$

    M[x,ū]
  - Compute the *expected sufficient statistics* for each x,**u**

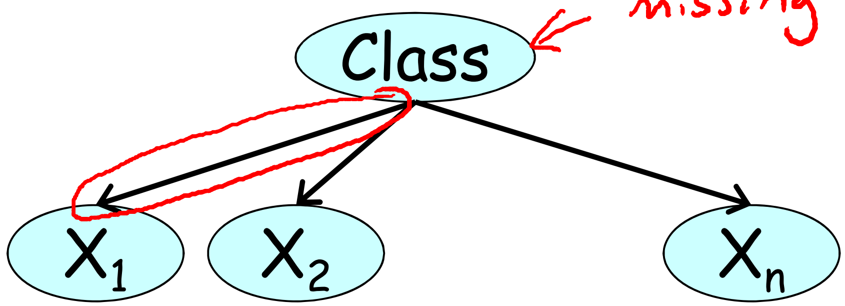
$$\overline{M}_{\theta^t}[x,\boldsymbol{u}] = \sum_{m=1}^{M} P(x,\boldsymbol{u} \mid \boldsymbol{d}[m],\theta^t)$$

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

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

Example: Bayesian Clustering

missing

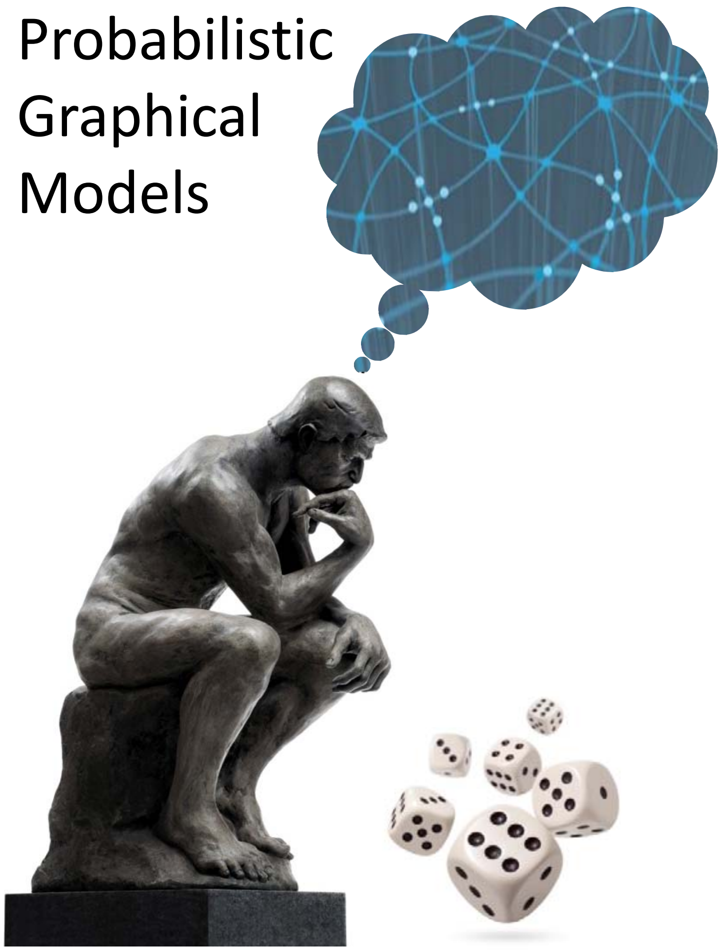


$$\bar{M}_{\boldsymbol{\theta}}[c] := \sum_m P(c \mid x_1[m], \ldots, x_n[m], \boldsymbol{\theta}^t) \qquad \theta_c^{t+1} = \frac{\bar{M}_{\boldsymbol{\theta}}[c]}{M}$$

$$\bar{M}_{\boldsymbol{\theta}}[x_i, c] := \sum_m P(c, x_i \mid x_1[m], \ldots, x_n[m], \boldsymbol{\theta}^t) \qquad \theta_{x_i|c}^{t+1} := \frac{\bar{M}_{\boldsymbol{\theta}}[x_i, c]}{\bar{M}_{\boldsymbol{\theta}}[c]}$$

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

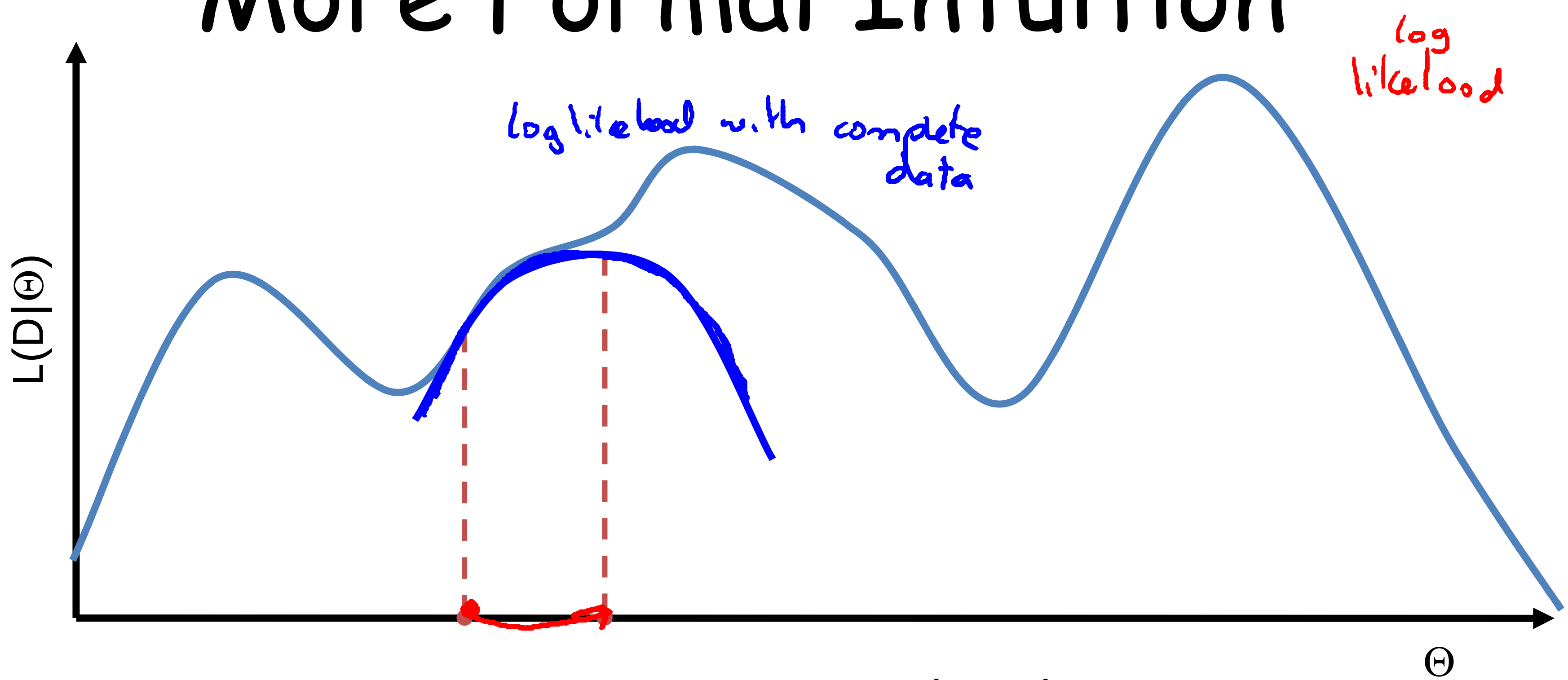


Learning

Incomplete Data

EM Analysis

More Formal Intuition



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

Daphne Koller

More Formal Intuition

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

assignment to H

log parameter

$$\ell(\boldsymbol{\theta} : \langle \boldsymbol{d}, \boldsymbol{h} \rangle) = \sum_{i=1}^{n} \sum_{(\boldsymbol{x}_i, \boldsymbol{u}_i) \in Val(X_i, \mathbf{Pa}_{X_i})} \mathbf{1}_{\langle \boldsymbol{d}, \boldsymbol{h} \rangle}[\boldsymbol{x}_i, \boldsymbol{u}_i] \log \theta_{\boldsymbol{x}_i | \boldsymbol{u}_i}$$

$$\boldsymbol{E}_{Q(\boldsymbol{H})}[\ell(\boldsymbol{\theta} : \langle \boldsymbol{d}, \boldsymbol{H} \rangle)] = \sum_{i=1}^{n} \sum_{(x_i, \boldsymbol{u}_i) \in Val(X_i, \mathbf{Pa}_{X_i})} \boldsymbol{E}_{Q(\boldsymbol{H})}\left[\mathbf{1}_{\langle \boldsymbol{d}, \boldsymbol{H} \rangle}[x_i, \boldsymbol{u}_i]\right] \log \theta_{x_i | \boldsymbol{u}_i}$$

$$= \sum_{i=1}^{n} \sum_{(\boldsymbol{x}_i, \boldsymbol{u}_i) \in Val(X_i, \mathbf{Pa}_{X_i})} Q(\boldsymbol{x}_i, \boldsymbol{u}_i) \log \theta_{\boldsymbol{x}_i | \boldsymbol{u}_i}$$

More Formal Intuition

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

$$Q^t_m(\boldsymbol{H}[m]) = P(\boldsymbol{H}[m] \mid \boldsymbol{d}[m], \boldsymbol{\theta}^t)$$

$$\sum_{m=1}^{M} \boldsymbol{E}_{Q^t_m(\boldsymbol{H}[m])}[\ell(\boldsymbol{\theta} : \langle \boldsymbol{d}[m], \boldsymbol{H}[m] \rangle)]$$

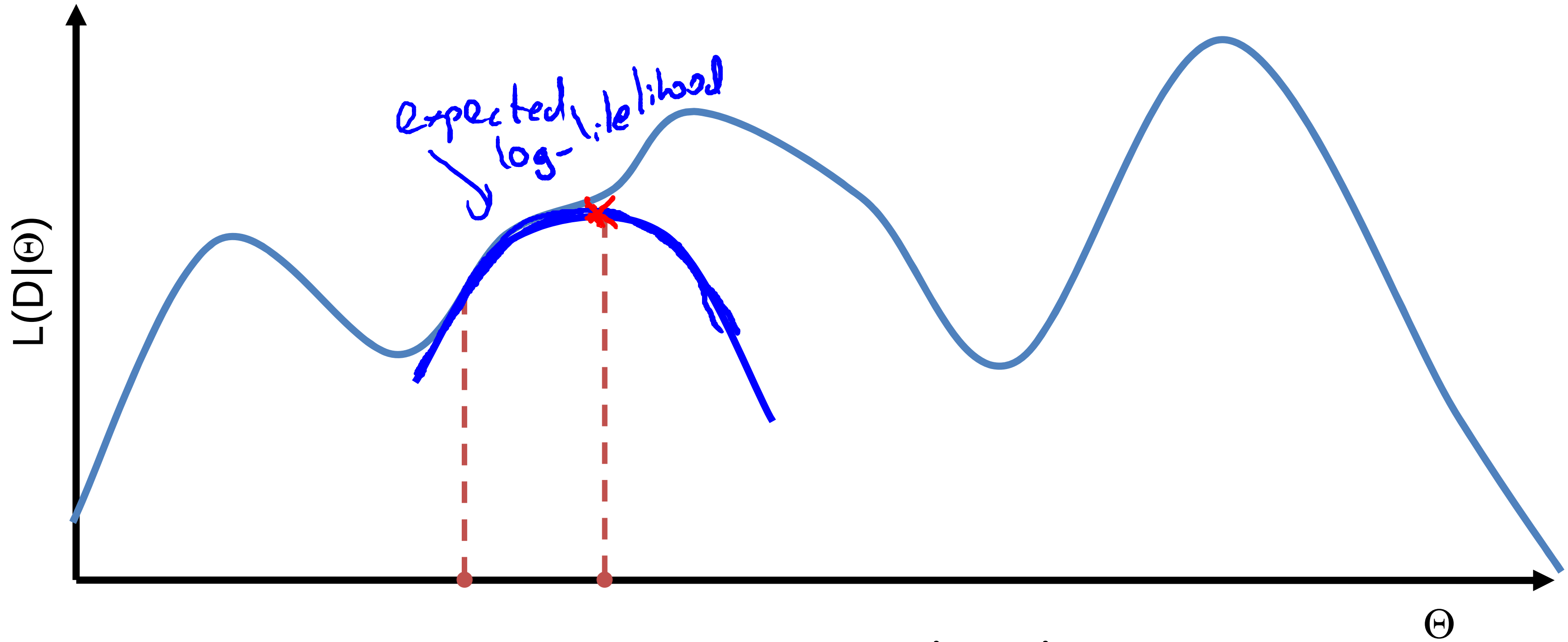
$$= \sum_{i=1}^{n} \sum_{(x_i, \boldsymbol{u}_i)} \left(\sum_{m=1}^{M} P(x_i, \boldsymbol{u}_i \mid \boldsymbol{d}[m], \boldsymbol{\theta}^t) \right) \log \theta_{x_i | \boldsymbol{u}_i}$$

$$= \sum_{i=1}^{n} \sum_{(x_i, \boldsymbol{u}_i)} \underbrace{\bar{M}_{\boldsymbol{\theta}^t}[x_i, \boldsymbol{u}_i]}_{\text{ESS}} \log \theta_{x_i | \boldsymbol{u}_i}$$

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

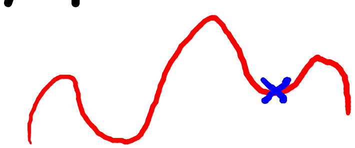
Daphne Koller

EM Guarantees

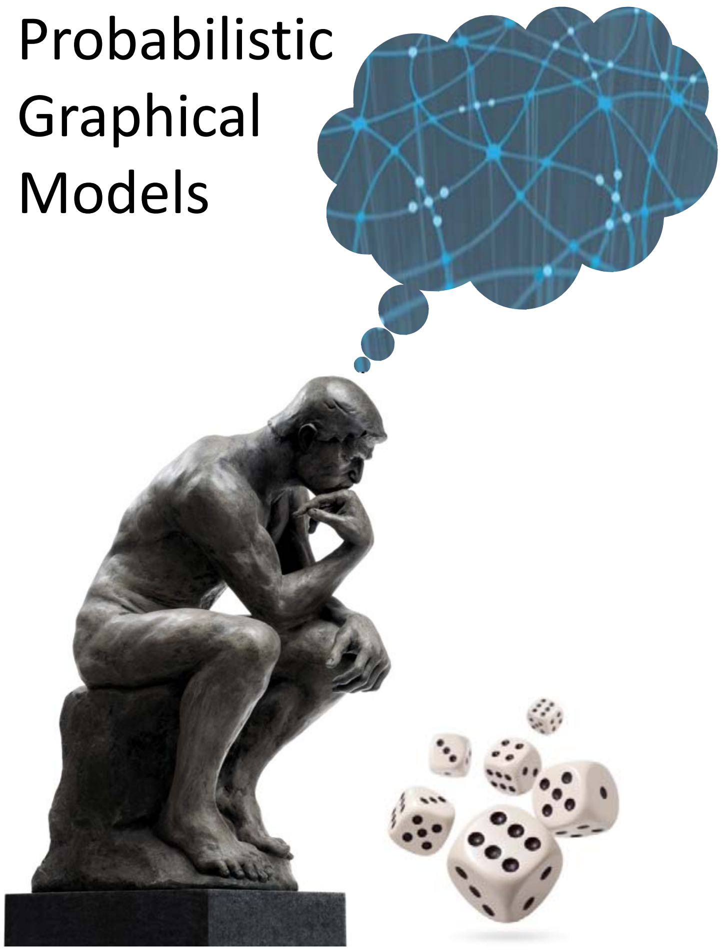
- $L(D : \theta^{t+1}) \geq L(D : \theta^{t})$
  - Each iteration improves the likelihood

gradient is zero

- If $\theta^{t+1} = \theta^{t}$, then θ^{t} is a stationary point of $L(D : \theta)$
  - Usually, this means a local maximum



Probabilistic
Graphical
Models

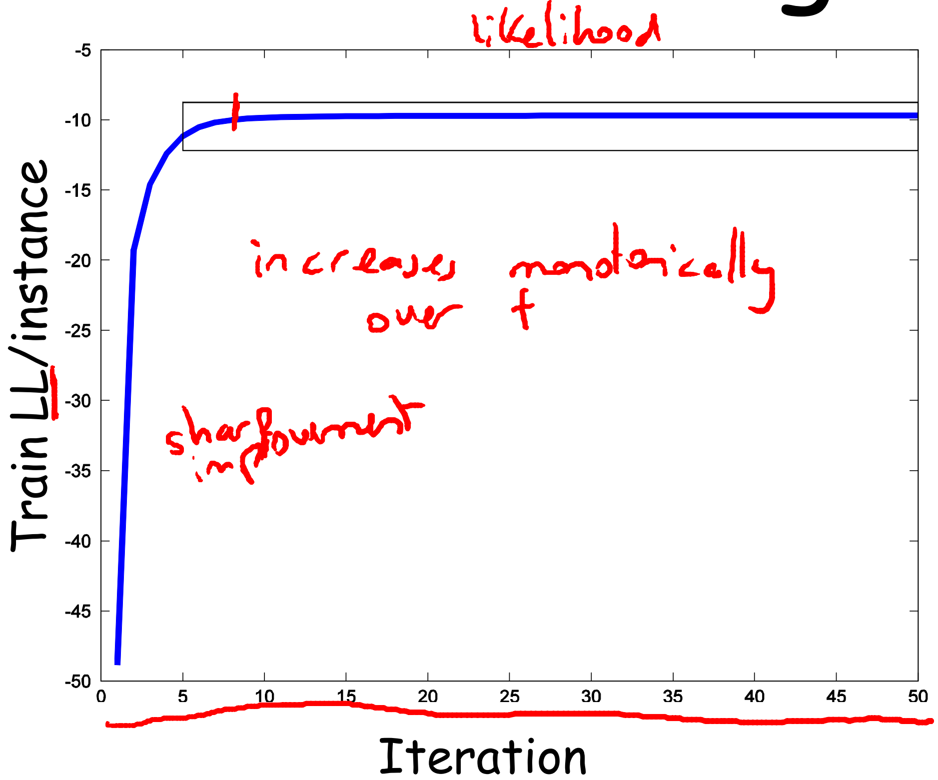


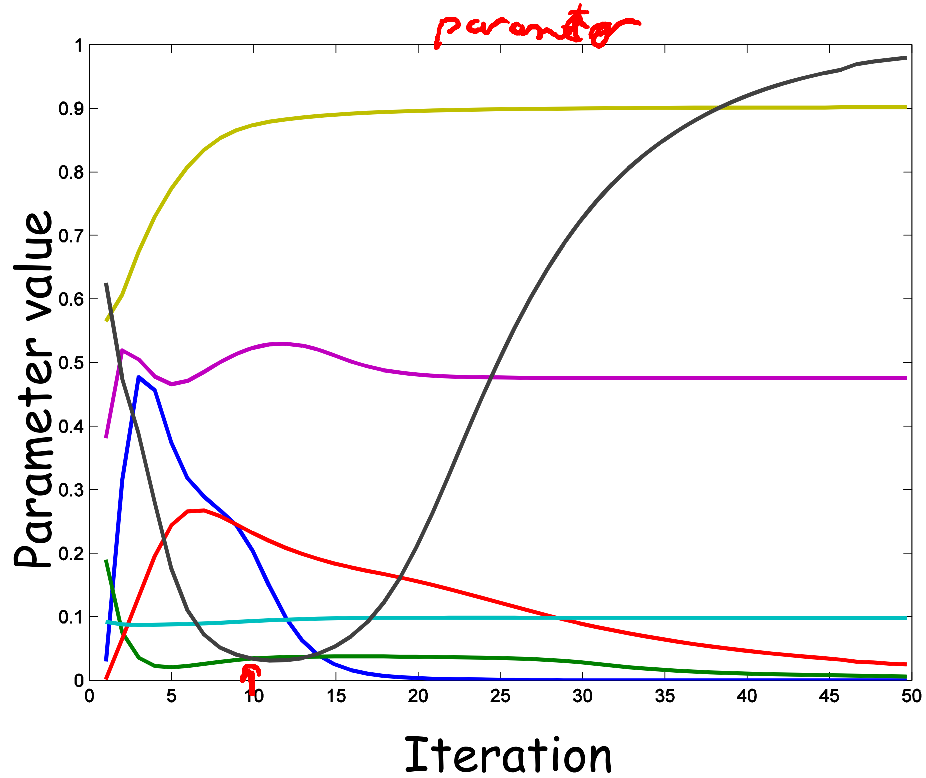
Learning

Incomplete Data

EM in Practice

EM Convergence in Practice

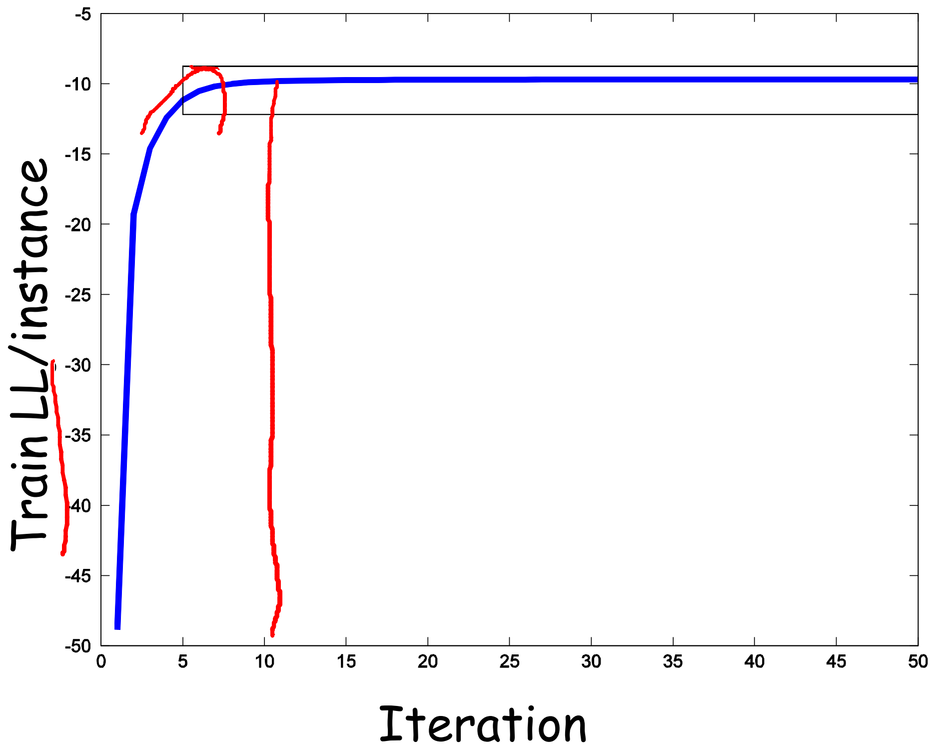


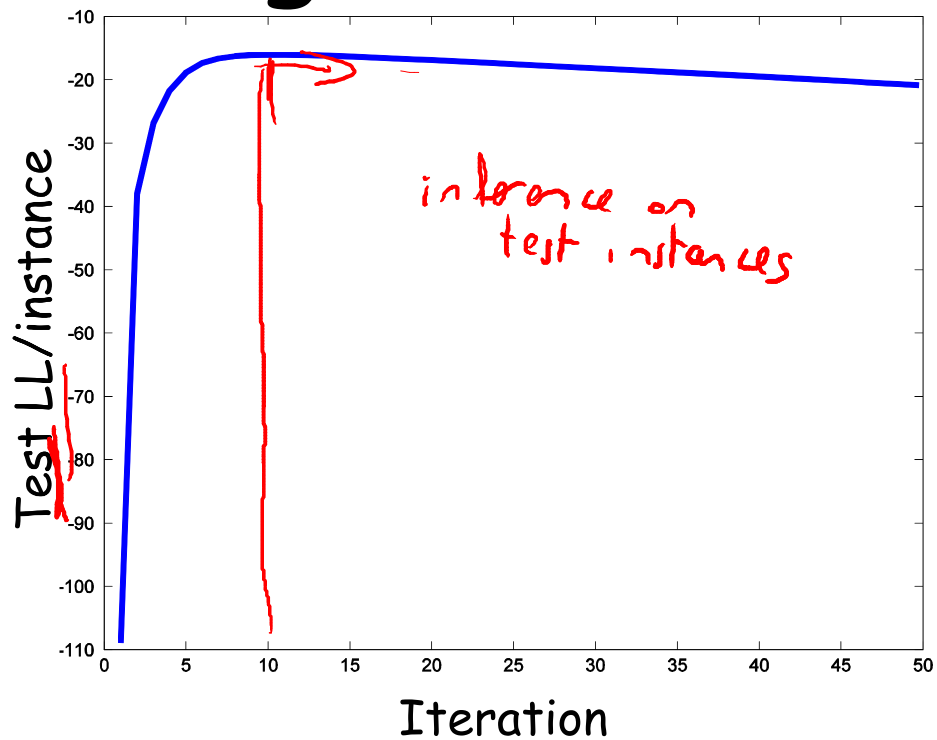


G. Elidan

Daphne Koller

Overfitting (numerical)

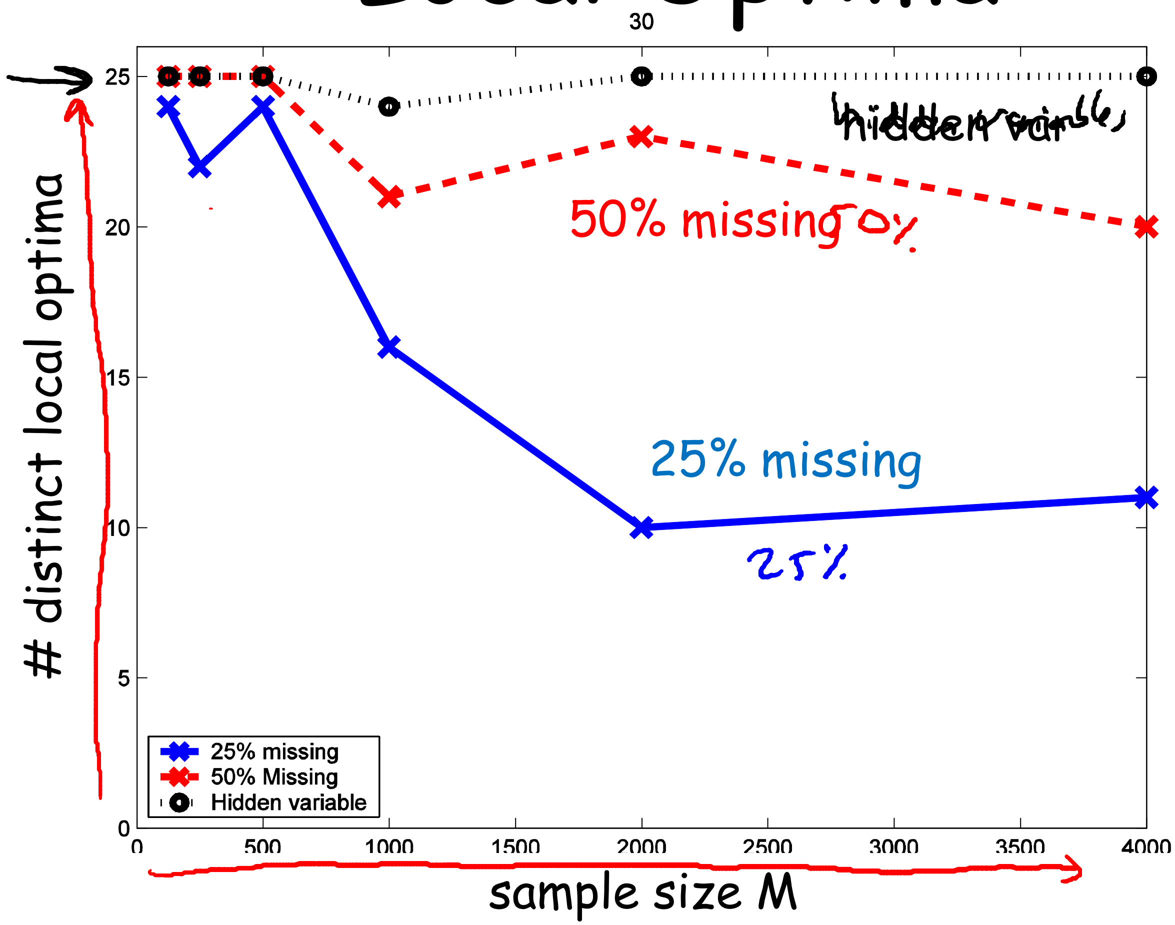




- Early stopping using cross validation
- Use MAP with parameter priors rather than MLE

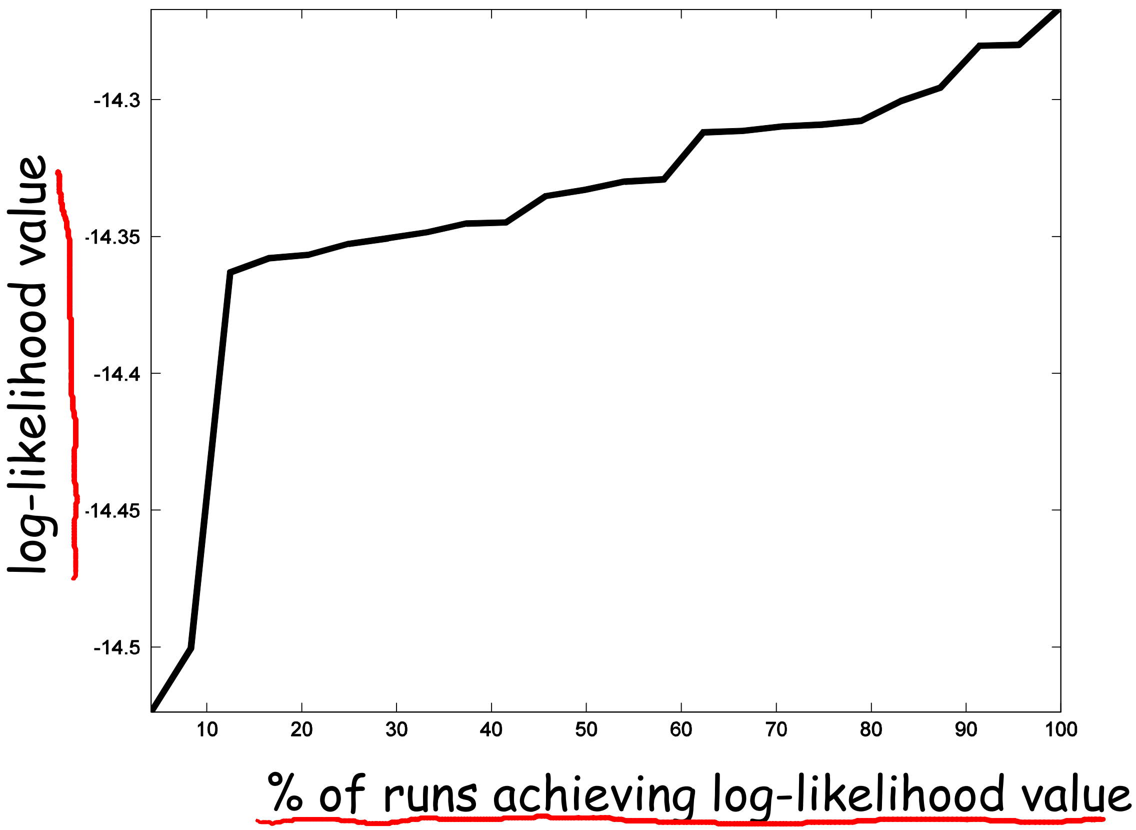
G. Elidan

Local Optima



G. Elidan

Significance of Local Optima



G. Elidan

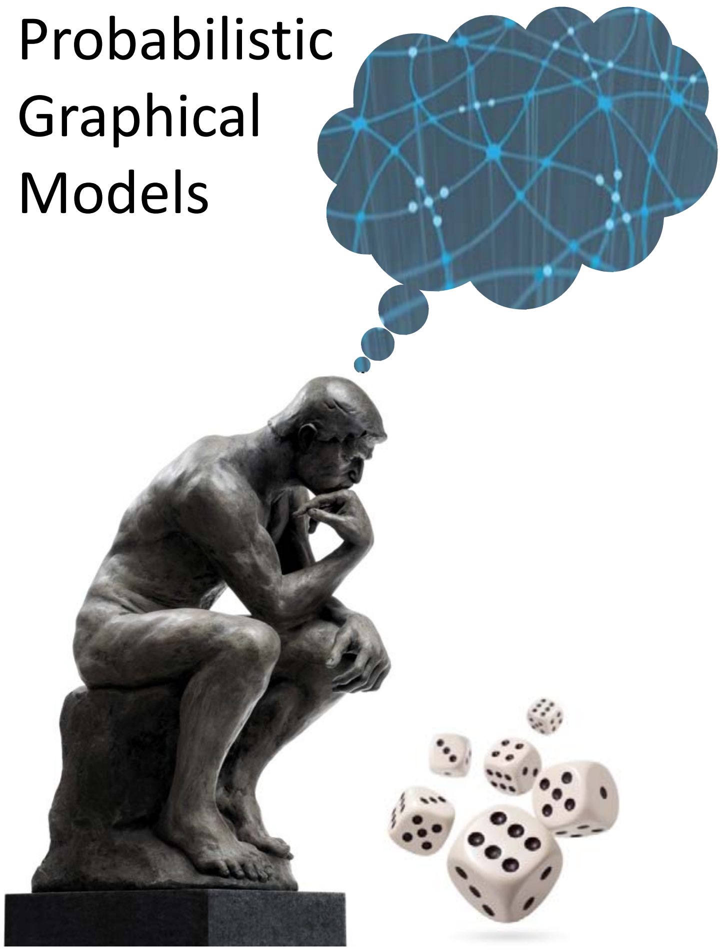
Initialization is Critical

- 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



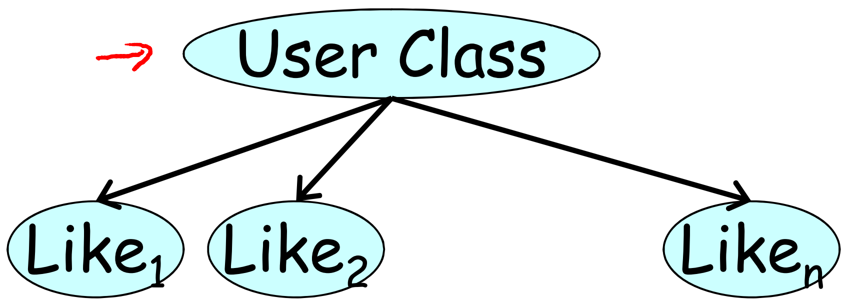
Learning

Incomplete Data

Learning with Latent Variables

Discovering User Clusters

class



features

segmenting users

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%):

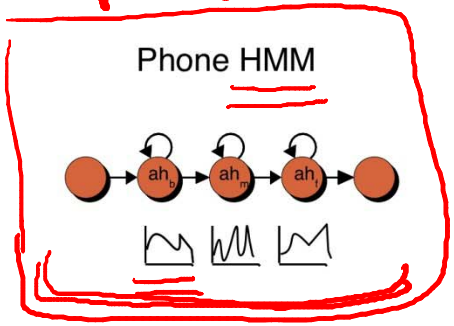
- The truth about what things cost
- Fuhrman Pleads Innocent To Perjury
- Real Astrology

Speech Recognition HMM

Lexicon

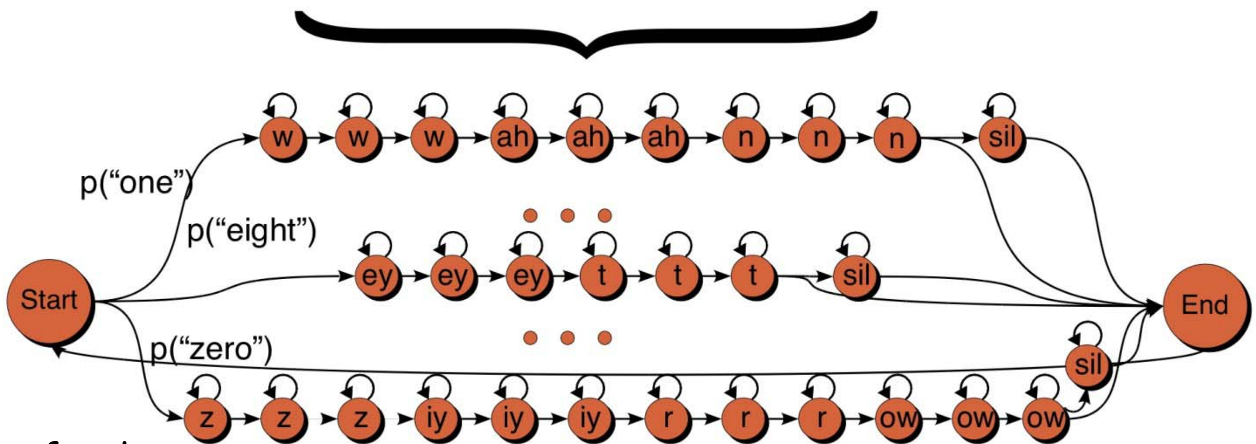
one	w ah n
two	t uw
three	th r iy
four	f ao r
f ve	f ay v
six	s ih k s
seven	s eh v ax n
eight	ey t
nine	n ay n
zero	z iy r ow

phone database



transition probabilistic

hidden variables



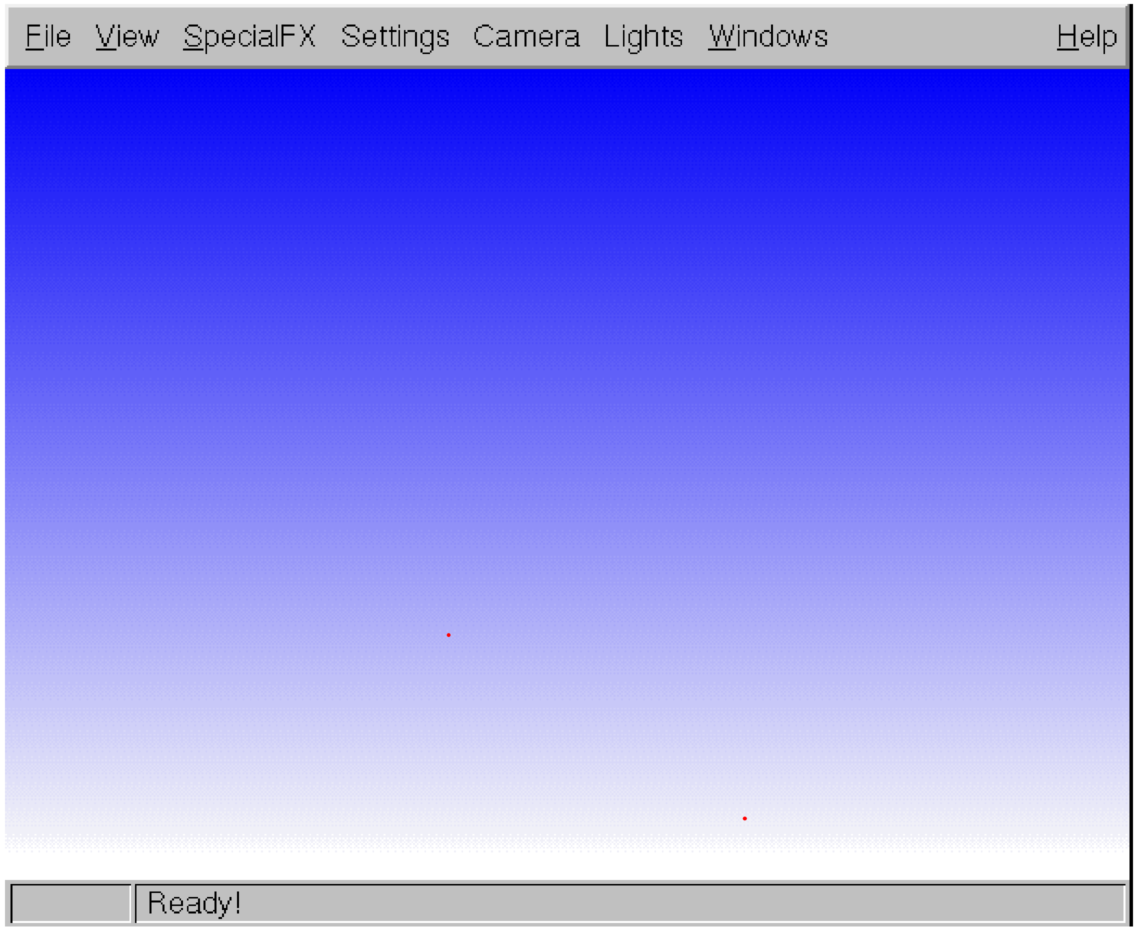
Dan Jurafsky, Stanford

3D Robot Mapping

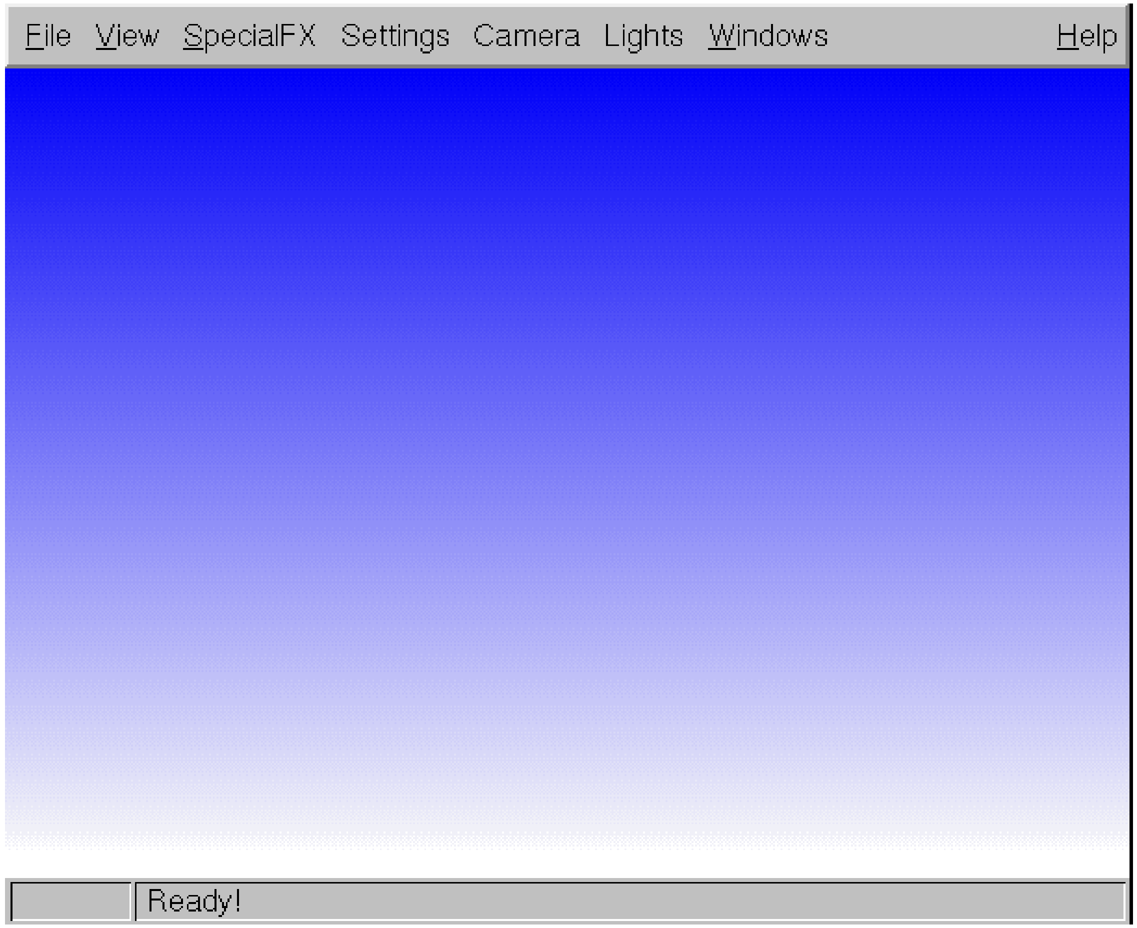
- Input: Point cloud from laser range finder obtained by moving robot
- Output: 3D planar map of environment

- 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



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

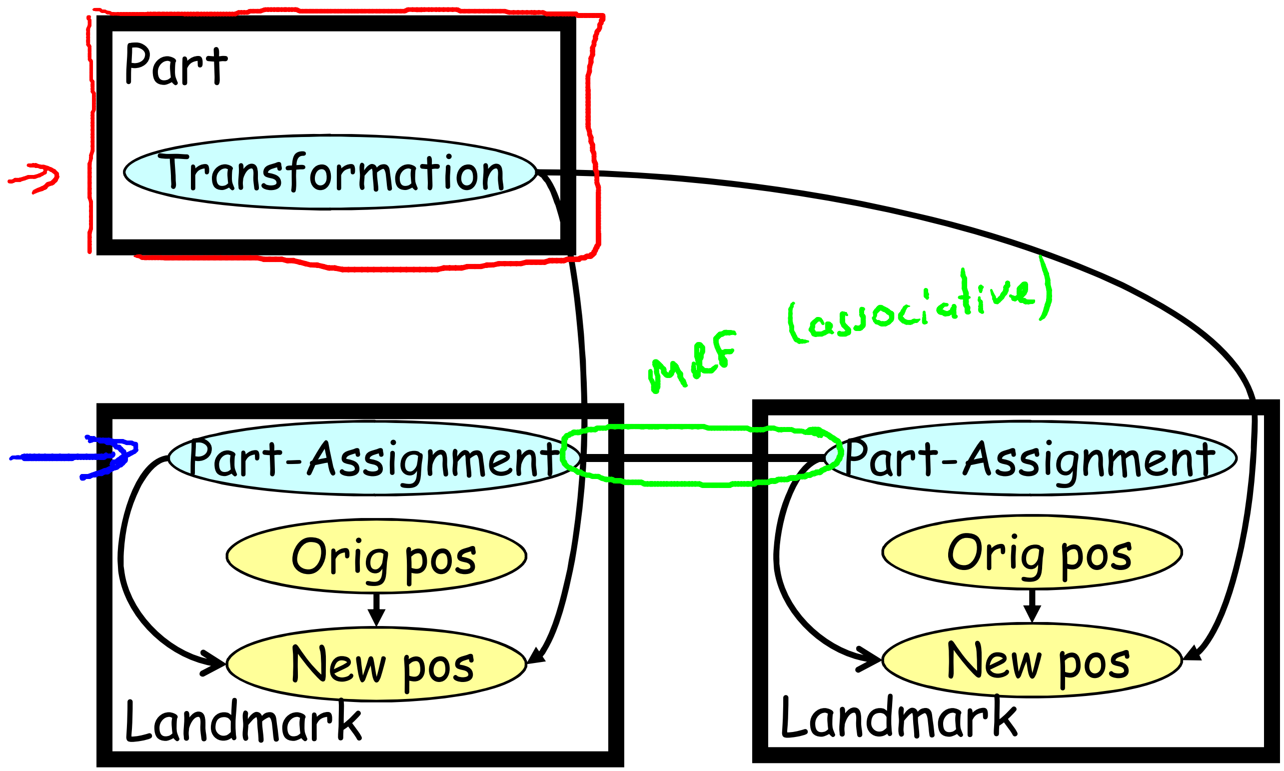


Thrun, Martin, Liu, Haehnel, Emery-Montemerlo, Chakrabarti, Burgard,
IEEE Transactions on Robotics, 2004

Body Parts from Point Cloud Scans

Anguelov, Koller, Pang, Srinivasan, Thrun UAI 2004

Collective Clustering Model



Anguelov, Koller, Pang, Srinivasan, Thrun UAI 2004

Anguelov, Koller, Pang, Srinivasan, Thrun UAI 2004

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

Graphical model

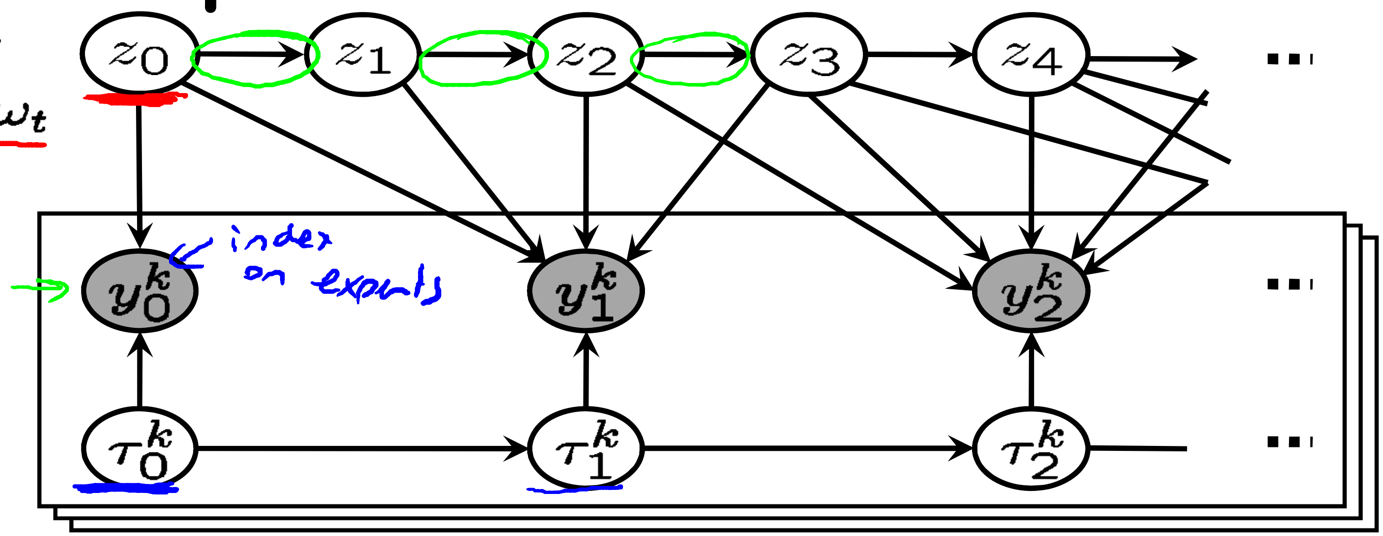
Intended trajectory

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

Expert demonstrations

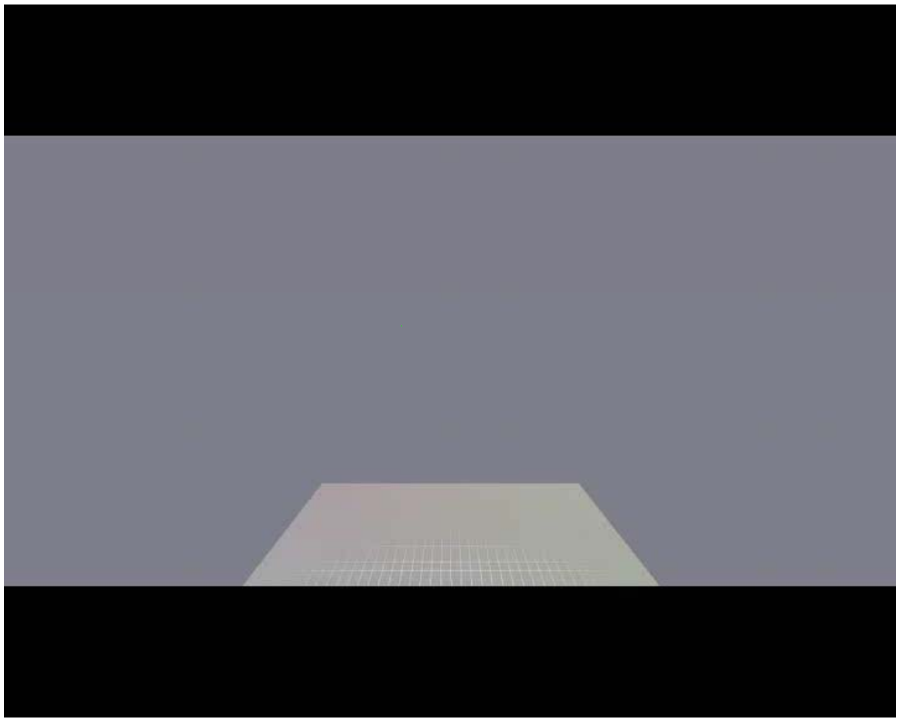
$$y_j = z_{\tau_j} + \nu_j$$

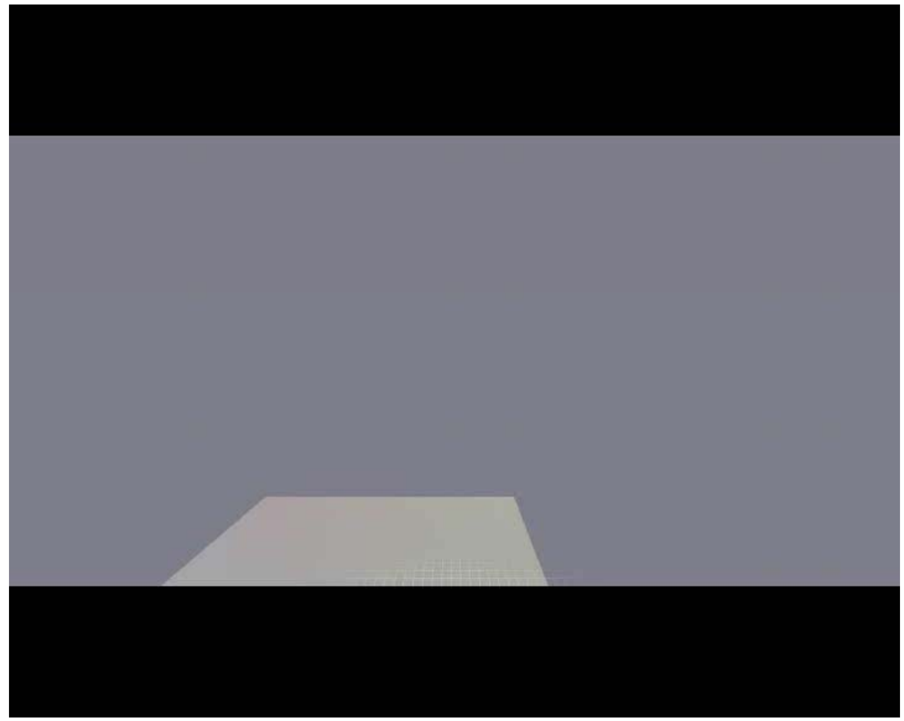
Time indices



Coates, Abbeel, Ng, ICML 2008

All Expert Demos





Coates, Abbeel, Ng, ICML 2008

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