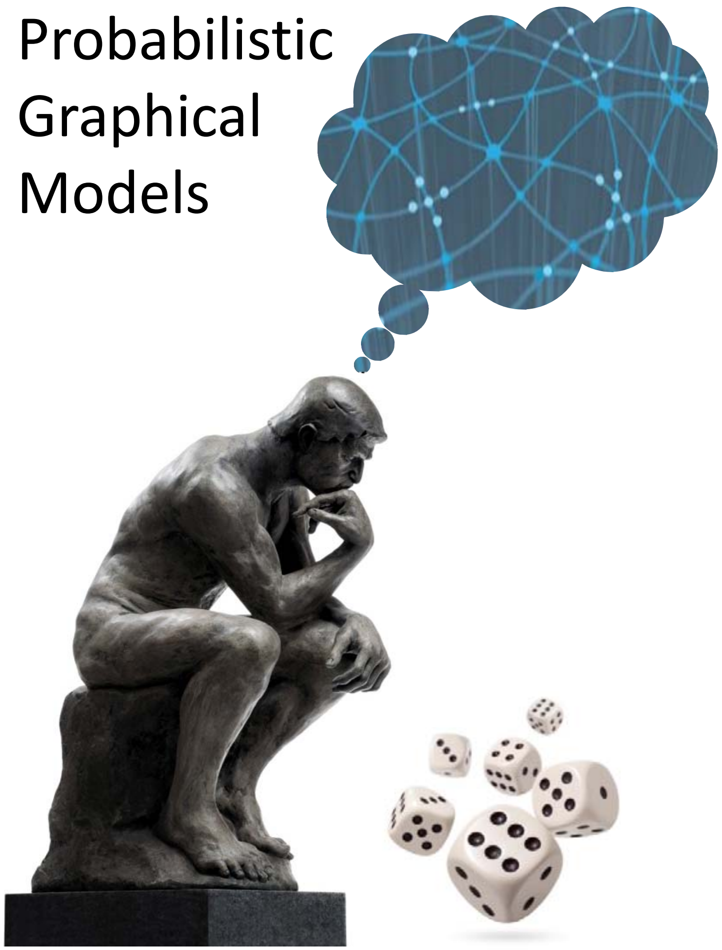
Probabilistic Graphical Models

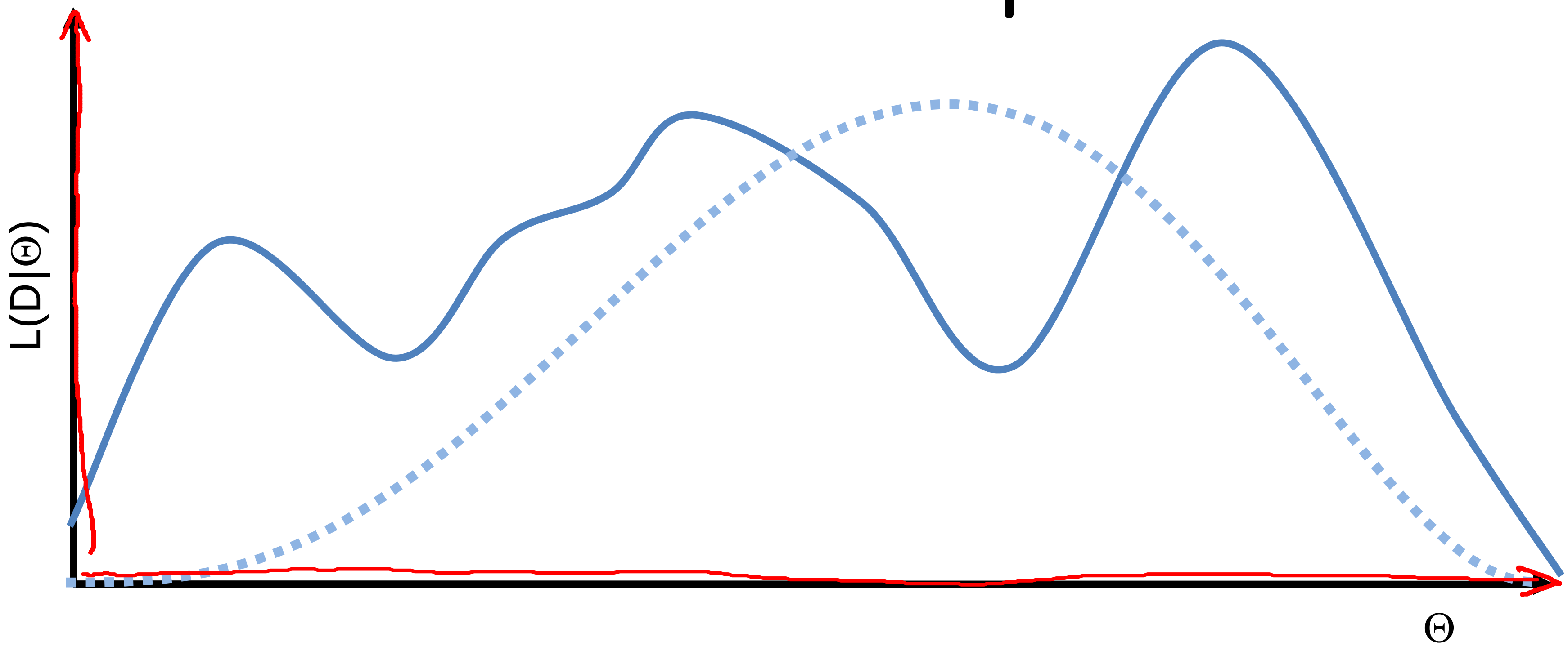


Learning

Incomplete Data

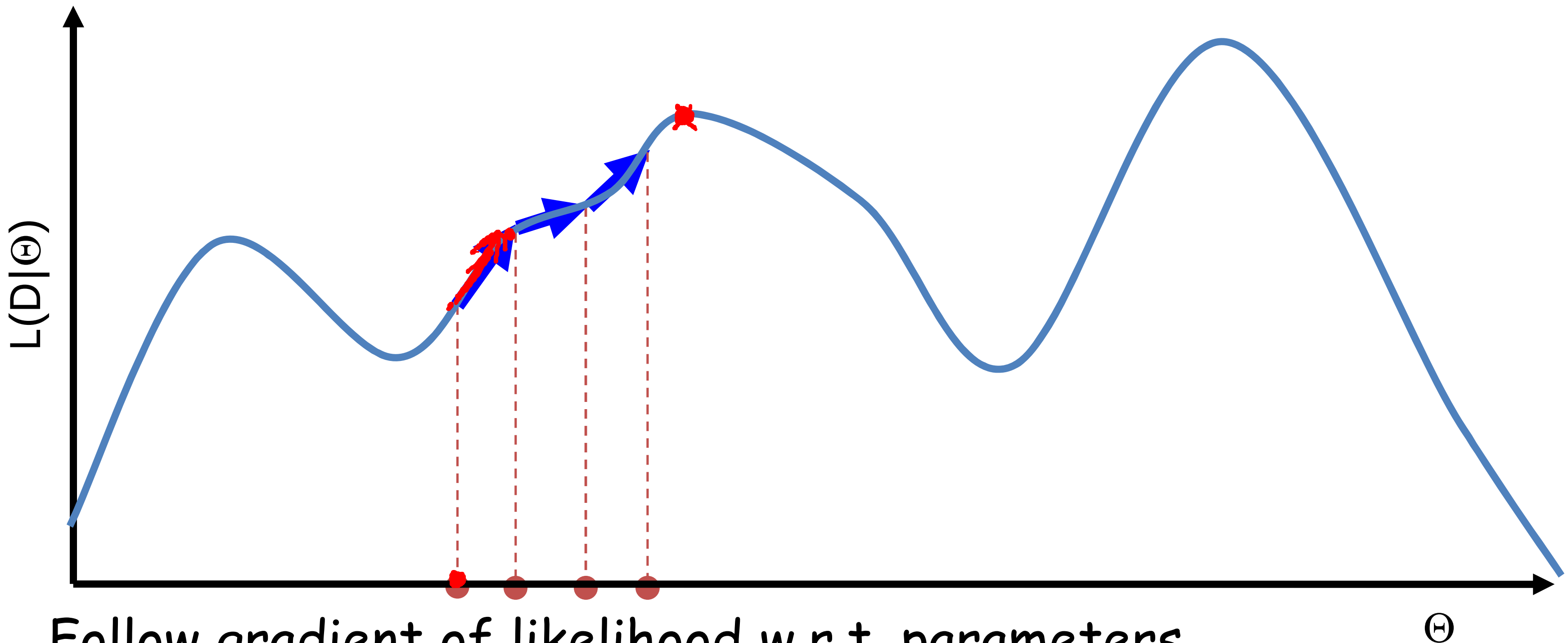
Likelihood Optimization Methods

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:

$$\frac{\partial \log P(D \mid \Theta)}{\partial \theta_{x_i|\mathrm{u}_i}} = \frac{1}{\theta_{x_i|\mathrm{u}_i}} \underbrace{\sum_m P(x_i, \mathrm{u}_i \mid \overbrace{\mathrm{d}[m]}^{\text{evidence in } m\text{th instance}}, \overbrace{\Theta}^{\text{current param value}})}_{\text{data instances } m}$$

- 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

Daphne Koller

Expectation Maximization (EM)

- Expectation (E-step):
  - For each data case **d**[m] and each family X,**U** compute

    soft completion

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

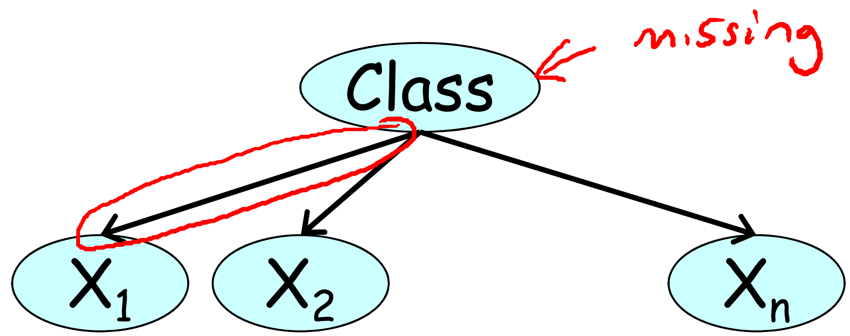
    $M[x,\bar{u}]$
  - Compute the *expected sufficient statistics* for each x,**u**

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

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

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

Example: Bayesian Clustering



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

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

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