

Probabilistic Graphical Models

Learning with partial observations

Siamak Ravanbakhsh

Winter 2021

Learning objectives

- different types of missing data
- learning with missing data and hidden vars:
 - directed models
 - undirected models
- develop an intuition for expectation maximization
 - variational interpretation

Two settings for partial observations

- missing data
 - each instance in \mathcal{D} is missing some values

Two settings for partial observations

- missing data
 - each instance in \mathcal{D} is missing some values
- hidden variables
 - variables that are never observed

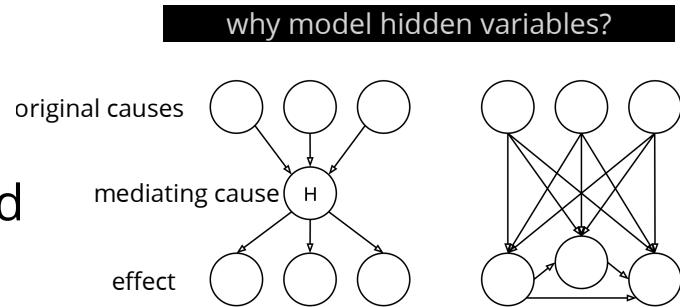


image credit: Murphy's book

Two settings for partial observations

- missing data
 - each instance in \mathcal{D} is missing some values
- hidden variables
 - variables that are never observed

latent variable models

- observations have common cause
- widely used in machine learning

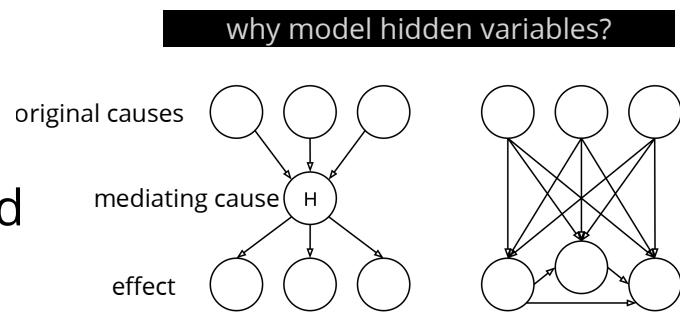
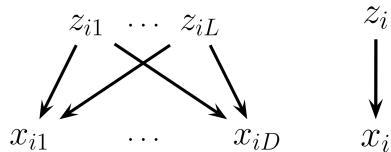
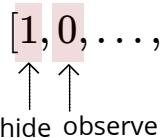


image credit: Murphy's book

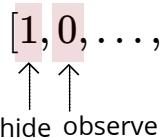
Missing data

observation mechanism:

- generate the data point $X = [X_1, \dots, X_D]$
- decide the values to observe $O_X = [1, 0, \dots, 0, 1]$


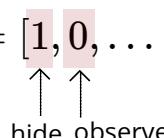
Missing data

observation mechanism:

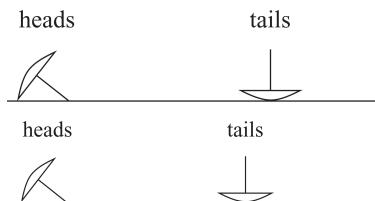
- generate the data point $X = [X_1, \dots, X_D]$
- decide the values to observe $O_X = [1, 0, \dots, 0, 1]$

hide observe
- observe X_o while X_h is missing ($X = [X_h; X_o]$)

Missing data

observation mechanism:

- generate the data point $X = [X_1, \dots, X_D]$
- decide the values to observe $O_X = [1, 0, \dots, 0, 1]$

- observe X_o while X_h is missing ($X = [X_h; X_o]$)

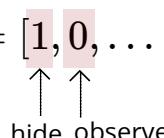
missing completely at random (MCAR) $P(X, O_X) = P(X)P(O_X)$



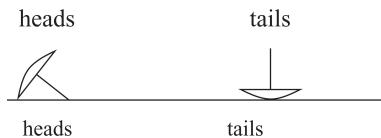
$$p_\theta(x) = \theta^x(1 - \theta)^{1-x} \text{ throw to generate}$$

Missing data

observation mechanism:

- generate the data point $X = [X_1, \dots, X_D]$
- decide the values to observe $O_X = [1, 0, \dots, 0, 1]$

- observe X_o while X_h is missing ($X = [X_h; X_o]$)

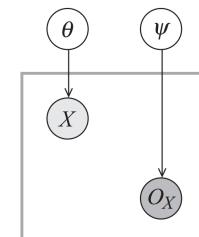
missing completely at random (MCAR) $P(X, O_X) = P(X)P(O_X)$



$$p_\theta(x) = \theta^x(1 - \theta)^{1-x} \text{ throw to generate}$$



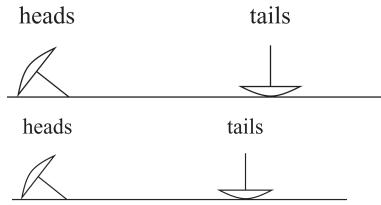
$$p_\psi(o) = \psi^o(1 - \psi)^{1-o} \text{ throw to decide show/hide}$$



Learning with MCAR

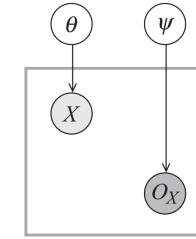
missing completely at random (MCAR)

$$P(X, O) = P(X)P(O)$$



$$p_\theta(x) = \theta^x(1 - \theta)^{1-x} \text{ throw to generate}$$

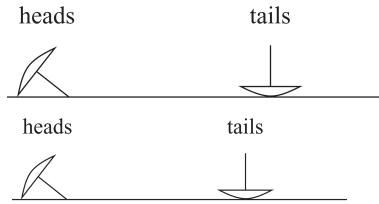
$$p_\psi(o) = \psi^o(1 - \psi)^{1-o} \text{ throw to decide show/hide}$$



Learning with MCAR

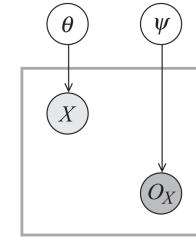
missing completely at random (MCAR)

$$P(X, O) = P(X)P(O)$$



$$p_\theta(x) = \theta^x(1 - \theta)^{1-x} \text{ throw to generate}$$

$$p_\psi(o) = \psi^o(1 - \psi)^{1-o} \text{ throw to decide show/hide}$$

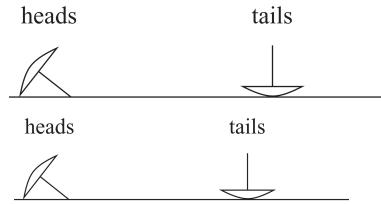


objective: learn a model for X , from the data $\mathcal{D} = \{x_o^{(1)}, \dots, x_o^{(M)}\}$
each x_o may include values for a different subset of vars.

Learning with MCAR

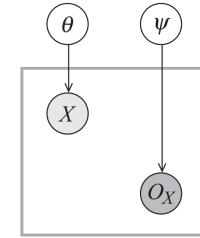
missing completely at random (MCAR)

$$P(X, O) = P(X)P(O)$$



$$p_\theta(x) = \theta^x(1 - \theta)^{1-x} \text{ throw to generate}$$

$$p_\psi(o) = \psi^o(1 - \psi)^{1-o} \text{ throw to decide show/hide}$$



objective: learn a model for X , from the data $\mathcal{D} = \{x_o^{(1)}, \dots, x_o^{(M)}\}$
each x_o may include values for a different subset of vars.

since $P(X, O) = P(X)P(O)$, we can ignore the obs. patterns

optimize log **marginal** likelihood $\ell(\mathcal{D}, \theta) = \sum_{x_o \in \mathcal{D}} \log \sum_{x_h} p_\theta(\textcolor{red}{x}_o, x_h)$

A more general criteria

missing at random (MAR) $O_X \perp X_h | X_o$

if there is information about the obs. pattern O_X in X_h
then it is also in observations X_o

A more general criteria

missing at random (MAR) $O_X \perp X_h | X_o$

if there is information about the obs. pattern O_X in X_h
then it is also in observations X_o

example

throw the thumb-tack twice $X = [X_1, X_2]$
if $X_2 = 1$ hide X_1
otherwise show X_1



missing at random

missing completely at random

A more general criteria

missing at random (MAR) $O_X \perp X_h | X_o$

if there is information about the obs. pattern O_X in X_h
then it is also in observations X_o

example

throw the thumb-tack twice $X = [X_1, X_2]$
if $X_2 = 1$ hide X_1
otherwise show X_1



missing at random

missing completely at random

no "extra" information in the **obs. pattern**, therefore ignore it

optimize: $\ell(\mathcal{D}, \theta) = \sum_{x_o \in \mathcal{D}} \log \sum_{x_h} p_\theta(x_o, x_h)$

marginal **Likelihood function**
for partial observations

- **fully observed** data:



- **directed**: log-likelihood decomposes
- **undirected**: does not decompose, but it is concave

marginal **Likelihood function** *for partial observations*

- **fully observed** data:



- **directed**: log-likelihood decomposes
- **undirected**: does not decompose, but it is concave

- **partially observed**:



- log-marginal likelihood does not decompose
- not concave anymore

marginal Likelihood function *for partial observations*

- fully observed data:



- **directed**: log-likelihood decomposes
- **undirected**: does not decompose, but it is concave

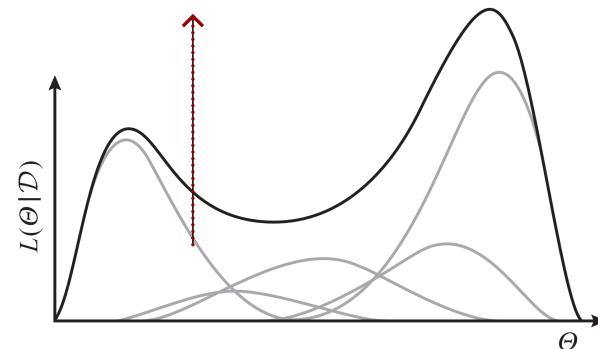
- partially observed:



- log-marginal likelihood does not decompose
- not concave anymore

$$\ell(\mathcal{D}, \theta) = \sum_{x_o \in \mathcal{D}} \log \sum_{x_h} p_\theta(x_o, x_h)$$

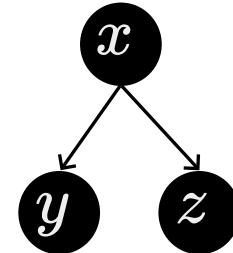
likelihood for a single assignment x_h
to the latent vars $p_\theta(x_o, x_h)$



marginal Likelihood function: example for a directed model

fully observed case decomposes:

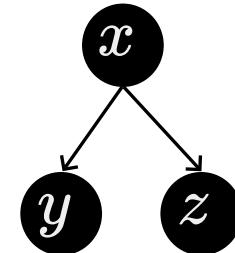
$$\begin{aligned}\ell(D, \theta) &= \sum_{x,y,z \in \mathcal{D}} \log p_\theta(x, y, z) \\ &= \sum_x \log p_{\theta_X}(x) + \sum_{x,y} \log p_{\theta_{Y|X}}(y|x) + \sum_{x,z} \log p_{\theta_{Z|X}}(z|x)\end{aligned}$$



marginal Likelihood function: example for a directed model

fully observed case decomposes:

$$\begin{aligned}\ell(D, \theta) &= \sum_{x,y,z \in \mathcal{D}} \log p_\theta(x, y, z) \\ &= \sum_x \log p_{\theta_X}(x) + \sum_{x,y} \log p_{\theta_{Y|X}}(y|x) + \sum_{x,z} \log p_{\theta_{Z|X}}(z|x)\end{aligned}$$



x is always missing (in a **latent variable model**)

$$\ell(D, \theta) = \sum_{y,z \in \mathcal{D}} \log \sum_x p_{\theta_X}(x) p_{\theta_{Y|X}}(y|x) p_{\theta_{Z|X}}(z|x)$$

cannot decompose the log marginal likelihood

Parameter learning with missing data

Directed models:

option 1: obtain the gradient of marginal likelihood

option 2: expectation maximization (EM)

- variational interpretation

Parameter learning with missing data

Directed models:

option 1: obtain the gradient of marginal likelihood

option 2: expectation maximization (EM)

- variational interpretation

undirected models:

obtain the gradient of marginal likelihood

- EM is not a good option here

Parameter learning with missing data

Directed models:

option 1: obtain the gradient of marginal likelihood

option 2: expectation maximization (EM)

- variational interpretation

undirected models:

obtain the gradient of marginal likelihood

- EM is not a good option here



all of these options
need **inference** for each step of
learning

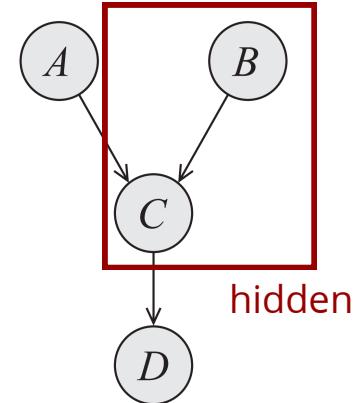
Gradient of the marginal likelihood

(directed models)

example

log marginal likelihood:

$$\ell(\mathcal{D}) = \sum_{(a,d) \in \mathcal{D}} \log \sum_{\textcolor{red}{b,c}} p(a)p(\textcolor{red}{b})p(\textcolor{red}{c}|a,\textcolor{red}{b})p(d|\textcolor{red}{c})$$



Gradient of the marginal likelihood

(directed models)

example

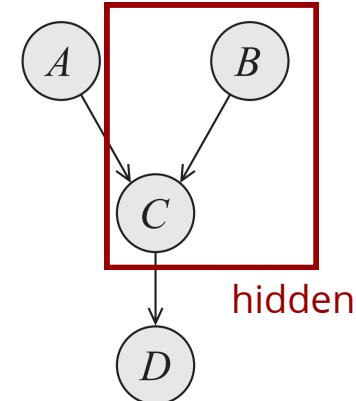
log marginal likelihood:

$$\ell(\mathcal{D}) = \sum_{(a,d) \in \mathcal{D}} \log \sum_{b,c} p(a)p(b)p(c|a,b)p(d|c)$$

take the derivative:

$$\frac{\partial}{\partial p(d'|c')} \ell(\mathcal{D}) = \frac{1}{p(d'|c')} \sum_{(a,d) \in \mathcal{D}} p(d', c' | a, d)$$

.....
need inference for this



Gradient of the marginal likelihood

(directed models)

example

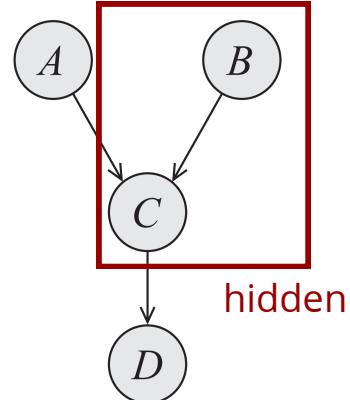
log marginal likelihood:

$$\ell(\mathcal{D}) = \sum_{(a,d) \in \mathcal{D}} \log \sum_{b,c} p(a)p(b)p(c|a,b)p(d|c)$$

take the derivative:

$$\frac{\partial}{\partial p(d'|c')} \ell(\mathcal{D}) = \frac{1}{p(d'|c')} \sum_{(a,d) \in \mathcal{D}} p(d', c' | a, d)$$

need inference for this
what happens to this expression if every variable is observed?



Gradient of the marginal likelihood

(directed models)

for a Bayesian Network with CPT

$$\frac{\partial}{\partial p(x_i|pa_{x_i})} \ell(\mathcal{D}) = \frac{1}{p(x_i|pa_{x_i})} \sum_{\mathbf{x}_o \in \mathcal{D}} p(x_i, pa_{x_i} \mid \mathbf{x}_o)$$

some specific assignment
run inference for each observation

Gradient of the marginal likelihood

(directed models)

for a Bayesian Network with CPT

$$\frac{\partial}{\partial p(x_i|pa_{x_i})} \ell(\mathcal{D}) = \frac{1}{p(x_i|pa_{x_i})} \sum_{\mathbf{x}_o \in \mathcal{D}} p(x_i, pa_{x_i} \mid \mathbf{x}_o)$$

some specific assignment
run inference for each observation

a technical issue:

- gradient is always non-negative
 - no constraint of the form $\sum_x p(x|pa_x) = 1$
 - reparametrize (e.g., using softmax)
 - or use Lagrange multipliers

Gradient of the marginal likelihood

(directed models)

for a Bayesian Network with CPT

$$\frac{\partial}{\partial p(x_i|pa_{x_i})} \ell(\mathcal{D}) = \frac{1}{p(x_i|pa_{x_i})} \sum_{\mathbf{x}_o \in \mathcal{D}} p(x_i, pa_{x_i} \mid \mathbf{x}_o)$$

some specific assignment
run inference for each observation

a technical issue:

- gradient is always non-negative
 - no constraint of the form $\sum_x p(x|pa_x) = 1$
 - reparametrize (e.g., using softmax)
 - or use Lagrange multipliers

for other parametrizations (beyond simple CPTs) use the chain rule:

$$\frac{\partial}{\partial \theta} \ell(\mathcal{D}; \theta) = \sum_{(c', d')} \frac{\partial \ell(\mathcal{D})}{\partial p(d' | c')} \frac{\partial p(d' | c')}{\partial \theta}$$

example

Expectation Maximization

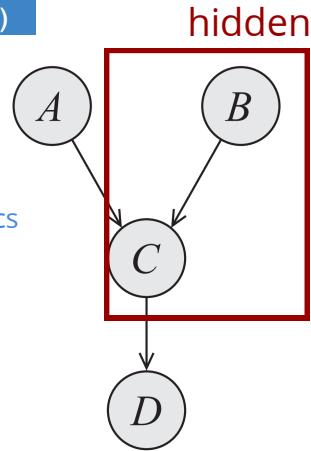
(directed models)

E-step:

for each $a, d \in \mathcal{D}$

use the current parameters θ to get the marginals

more generally: expected sufficient statistics
↓



example

Expectation Maximization

(directed models)

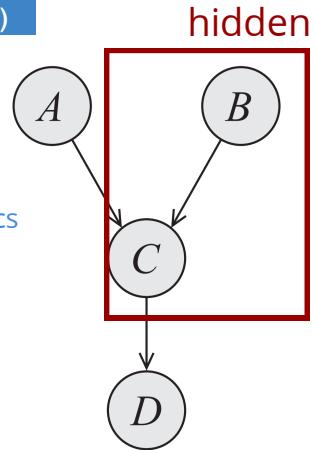
E-step:

for each $a, d \in \mathcal{D}$

use the current parameters θ to get the marginals

more generally: expected sufficient statistics

$p_{\theta, \mathcal{D}}(B), p_{\theta, \mathcal{D}}(A), p_{\theta, \mathcal{D}}(C), p_{\theta, \mathcal{D}}(A, B, C), p_{\theta, \mathcal{D}}(D, C)$



example

Expectation Maximization

(directed models)

hidden

E-step:

for each $a, d \in \mathcal{D}$

use the current parameters θ to get the marginals

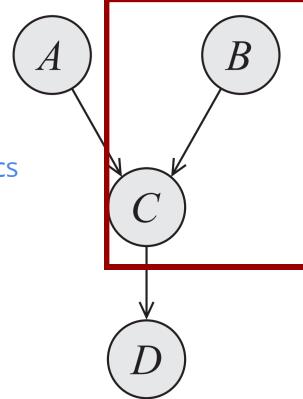
more generally: expected sufficient statistics

$$p_{\theta, \mathcal{D}}(B), p_{\theta, \mathcal{D}}(A), p_{\theta, \mathcal{D}}(C), p_{\theta, \mathcal{D}}(A, B, C), p_{\theta, \mathcal{D}}(D, C)$$

$$p_{\theta, \mathcal{D}}(C = c', D = d') = \frac{1}{N} \sum_{(a, d) \in \mathcal{D}} p_{\theta}(c', d' | a, d)$$

nonzero for $d' = d$

in general we need inference to estimate this sufficient statistics



example

Expectation Maximization

(directed models)

hidden

E-step:

for each $a, d \in \mathcal{D}$

use the current parameters θ to get the marginals

more generally: expected sufficient statistics

$$p_{\theta, \mathcal{D}}(B), p_{\theta, \mathcal{D}}(A), p_{\theta, \mathcal{D}}(C), p_{\theta, \mathcal{D}}(A, B, C), p_{\theta, \mathcal{D}}(D, C)$$

$$p_{\theta, \mathcal{D}}(C = c', D = d') = \frac{1}{N} \sum_{(a, d) \in \mathcal{D}} p_{\theta}(c', d' | a, d)$$

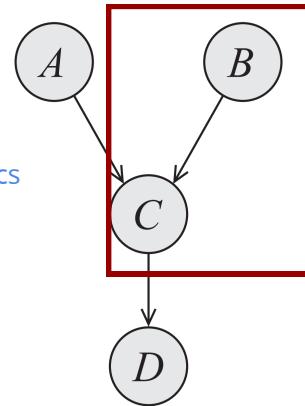
nonzero for $d' = d$

in general we need inference to estimate this sufficient statistics

M-step:

use the marginals (similar to completely observed data) to learn θ

expected sufficient statistics



example

Expectation Maximization

(directed models)

hidden

E-step:

for each $a, d \in \mathcal{D}$

use the current parameters θ to get the marginals

more generally: expected sufficient statistics

$$p_{\theta, \mathcal{D}}(B), p_{\theta, \mathcal{D}}(A), p_{\theta, \mathcal{D}}(C), p_{\theta, \mathcal{D}}(A, B, C), p_{\theta, \mathcal{D}}(D, C)$$

$$p_{\theta, \mathcal{D}}(C = c', D = d') = \frac{1}{N} \sum_{(a, d) \in \mathcal{D}} p_{\theta}(c', d' | a, d)$$

nonzero for $d' = d$

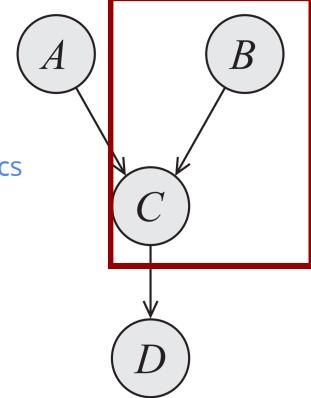
in general we need inference to estimate this sufficient statistics

M-step:

use the marginals (similar to completely observed data) to learn θ

expected sufficient statistics

E.g., update $\theta_{C|D}$ using $p_{\theta, \mathcal{D}}(C, D)$ and $p_{\theta, \mathcal{D}}(C)$ $\rightarrow \theta_{D|C}^{new} = \frac{p_{\theta, \mathcal{D}}(C, D)}{p_{\theta, \mathcal{D}}(C)}$



Expectation Maximization

(directed models)

for a Bayesian Network with CPT

E-step:

for each $\mathbf{x}_o \in \mathcal{D}$

use the current parameters θ to get the marginals

$$\{p_{\theta, \mathcal{D}}(X_i), p_{\theta, \mathcal{D}}(X_i, Pa_{X_i})\}$$

Expectation Maximization

(directed models)

for a Bayesian Network with CPT

E-step:

for each $\mathbf{x}_o \in \mathcal{D}$

use the current parameters θ to get the marginals

$$\{p_{\theta, \mathcal{D}}(X_i), p_{\theta, \mathcal{D}}(X_i, Pa_{X_i})\}$$

M-step:

use the marginals (similar to completely observed data) to learn θ^{new}

$$\theta_{X_i | Pa_{X_i}}^{new} = \frac{p_{\theta, \mathcal{D}}(X_i, Pa_{X_i})}{p_{\theta, \mathcal{D}}(Pa_{X_i})}$$

Expectation Maximization

(directed models)

for a Bayesian Network with CPT

E-step:

for each $\mathbf{x}_o \in \mathcal{D}$

use the current parameters θ to get the marginals

$$\{p_{\theta, \mathcal{D}}(X_i), p_{\theta, \mathcal{D}}(X_i, Pa_{X_i})\}$$

M-step:

use the marginals (similar to completely observed data) to learn θ^{new}

$$\theta_{X_i | Pa_{X_i}}^{new} = \frac{p_{\theta, \mathcal{D}}(X_i, Pa_{X_i})}{p_{\theta, \mathcal{D}}(Pa_{X_i})}$$

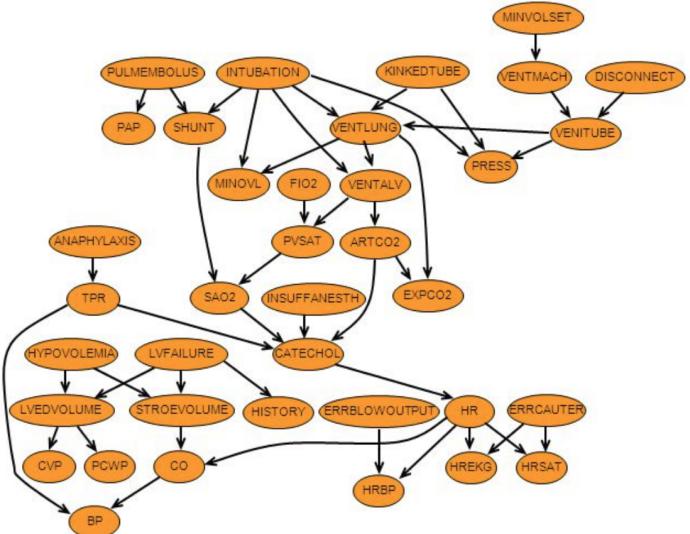
for **undirected models**: M-step is the expensive part

- perform E-step within each iteration of M-step: equivalent to gradient descent

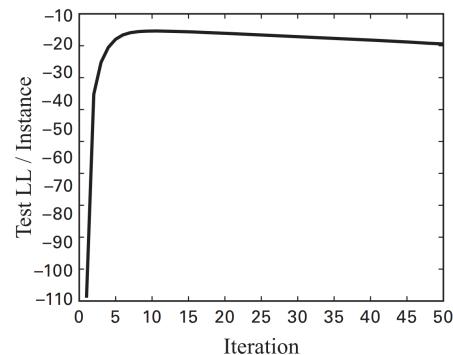
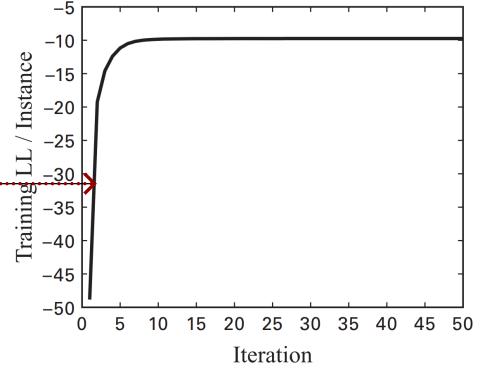
Expectation Maximization: example

alarm network

- 1000 training instances
- 50% of variables are observed (in each instance)



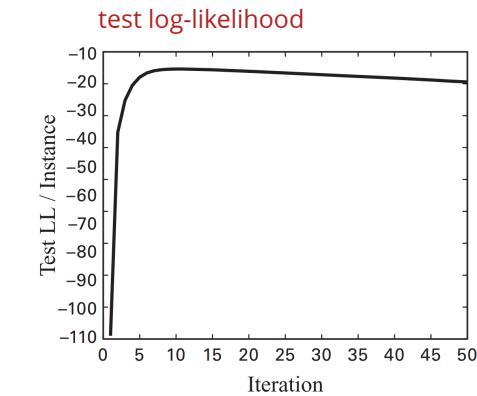
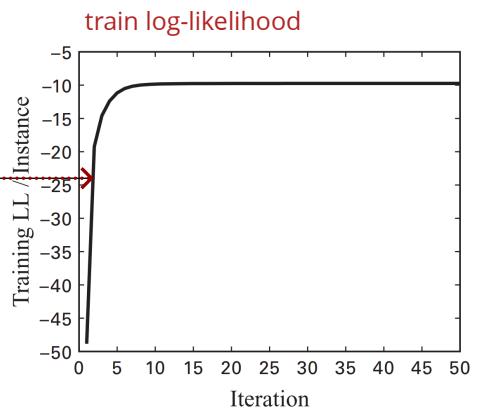
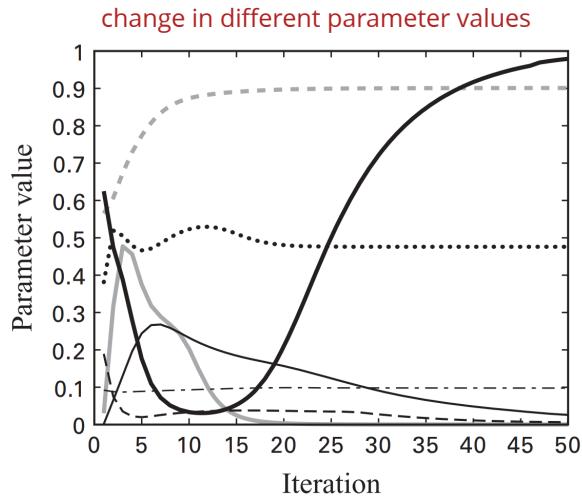
fast initial improvement



Expectation Maximization: example

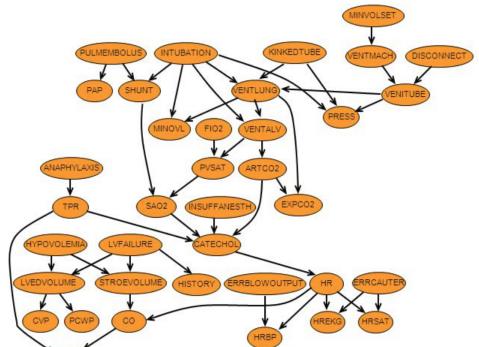
- 1000 training instances
- 50% of variables are observed (in each instance)

fast initial improvement

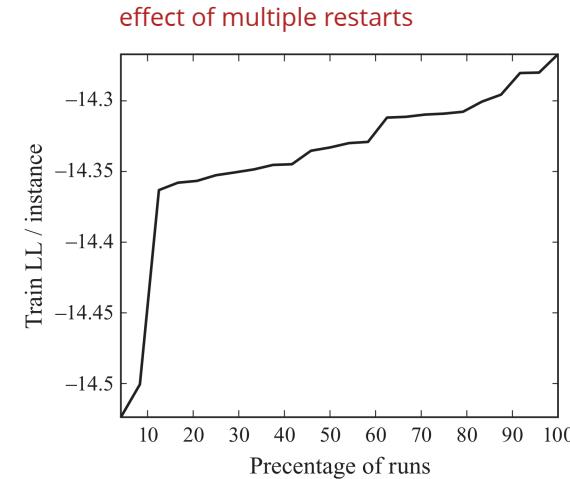
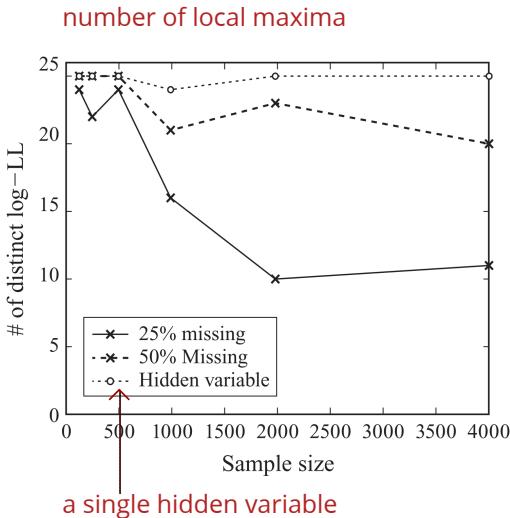


Expectation Maximization: example

local optima in EM:



alarm network



Expected log-likelihood

(directed models)

Original objective:

$$\ell(\mathcal{D}, \theta) = \sum_{\mathbf{x}_o \in \mathcal{D}} \log \frac{\sum_{\mathbf{x}_h} p_\theta(\mathbf{x}_o, \mathbf{x}_h)}{p_\theta(\mathbf{x}_o)}$$

Expected log-likelihood

(directed models)

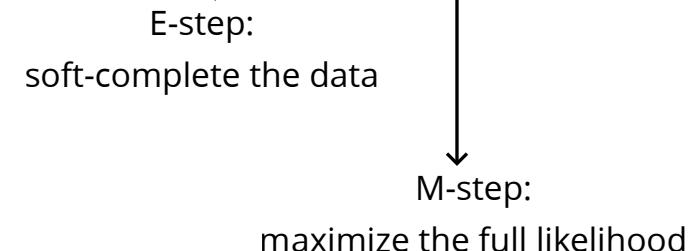
Original objective:

$$\ell(\mathcal{D}, \theta) = \sum_{\mathbf{x}_o \in \mathcal{D}} \log \frac{\sum_{\mathbf{x}_h} p_\theta(\mathbf{x}_o, \mathbf{x}_h)}{p_\theta(\mathbf{x}_o)}$$

EM iteration:

maximizes the expected log-likelihood

$$\sum_{\mathbf{x}_o \in \mathcal{D}} \mathbb{E}_{p_\theta(\mathbf{x}_h | \mathbf{x}_o)} [\log p_\theta(\mathbf{x}_o, \mathbf{x}_h)]$$



E-step:

soft-complete the data

M-step:

maximize the full likelihood

Expected log-likelihood

(directed models)

Original objective:

$$\ell(\mathcal{D}, \theta) = \sum_{\mathbf{x}_o \in \mathcal{D}} \log \frac{\sum_{\mathbf{x}_h} p_\theta(\mathbf{x}_o, \mathbf{x}_h)}{p_\theta(\mathbf{x}_o)}$$

EM iteration:

maximizes the expected log-likelihood

- how are these objectives related?
- any guarantees for EM?
- variational interpretation relates these two

$$\sum_{\mathbf{x}_o \in \mathcal{D}} \mathbb{E}_{p_\theta(\mathbf{x}_h | \mathbf{x}_o)} [\log p_\theta(\mathbf{x}_o, \mathbf{x}_h)]$$

E-step:

soft-complete the data

M-step:
maximize the full likelihood

Example: Gaussian mixture model

$$p(x; \pi) = \prod_k \pi_k^{\mathbb{I}(x=k)}$$

model parameters $\theta = [\pi, \{\mu_k, \Sigma_k\}]$

$$p(y|x; \{\mu_k, \Sigma_k\}) = \frac{1}{\sqrt{|2\pi\Sigma_x|}} \exp(-\frac{1}{2}(y - \mu_x)^T \Sigma_x^{-1} (y - \mu_x))$$

Example: Gaussian mixture model

$$p(x; \pi) = \prod_k \pi_k^{\mathbb{I}(x=k)}$$

model parameters $\theta = [\pi, \{\mu_k, \Sigma_k\}]$

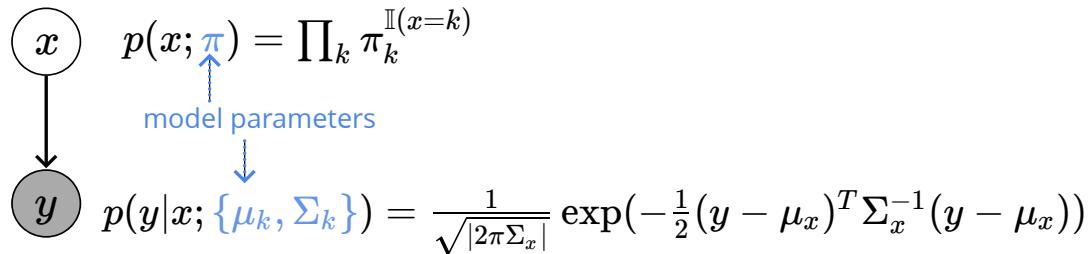
$$p(y|x; \{\mu_k, \Sigma_k\}) = \frac{1}{\sqrt{|2\pi\Sigma_x|}} \exp(-\frac{1}{2}(y - \mu_x)^T \Sigma_x^{-1} (y - \mu_x))$$

E-step: calculate $p(x|y)$ for each $y \in \mathcal{D}$

$$p(x = k|y) \propto p(x = k; \pi)p(y|x = k; \mu, \Sigma) = \pi_k \mathcal{N}(y; \mu_k, \Sigma_k)$$

- now we have "*probabilistically completed*" instances
- in **hard-EM** we find a single completion (rather than soft/probabilistic completion)
- next, update the parameters assuming complete data (easy for Bayes Nets)

Example: Gaussian mixture model



M-step: estimate $\pi, \mu_k, \Sigma_k \forall k$

$$\pi_k^{new} = \frac{1}{N} \sum_{y \in \mathcal{D}} \frac{p(x=k|y)}{\sum_{k'} p(x=k'|y)}$$

$$\mu_k = \frac{\sum_{y \in \mathcal{D}} p(x=k|y)y}{\sum_{y \in \mathcal{D}} p(x=k|y)} \quad \text{mean of a weighted set of instances}$$

$$\Sigma_k = \frac{\sum_{y \in \mathcal{D}} p(x=k|y)(y - \mu_k)(y - \mu_k)^T}{\sum_{y \in \mathcal{D}} p(x=k|y)} \quad \text{covariance of a weighted set of instances}$$

Example: Gaussian mixture model

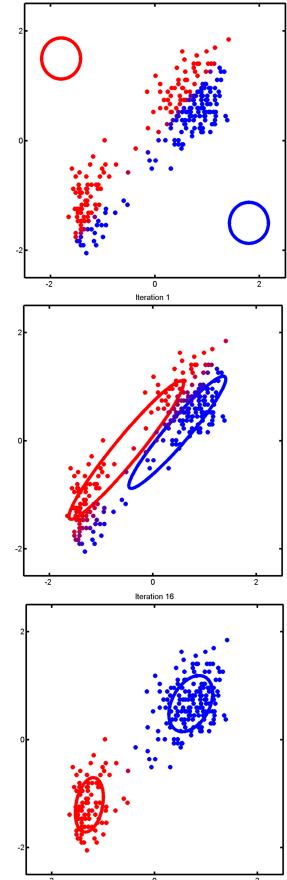
$$\begin{array}{c}
 \text{Diagram: } \\
 \begin{array}{ccc}
 \text{circle } x & \xrightarrow{\quad} & \text{circle } y \\
 \downarrow & & \downarrow \\
 \text{model parameters} & & \\
 \end{array}
 \end{array}
 \quad
 \begin{aligned}
 p(x; \pi) &= \prod_k \pi_k^{\mathbb{I}(x=k)} \\
 p(y|x; \{\mu_k, \Sigma_k\}) &= \frac{1}{\sqrt{|2\pi\Sigma_x|}} \exp\left(-\frac{1}{2}(y - \mu_x)^T \Sigma_x^{-1} (y - \mu_x)\right)
 \end{aligned}$$

M-step: estimate $\pi, \mu_k, \Sigma_k \forall k$

$$\pi_k^{new} = \frac{1}{N} \sum_{y \in \mathcal{D}} \frac{p(x=k|y)}{\sum_{k'} p(x=k'|y)}$$

$$\mu_k = \frac{\sum_{y \in \mathcal{D}} p(x=k|y)y}{\sum_{y \in \mathcal{D}} p(x=k|y)} \quad \text{mean of a weighted set of instances}$$

$$\Sigma_k = \frac{\sum_{y \in \mathcal{D}} p(x=k|y)(y - \mu_k)(y - \mu_k)^T}{\sum_{y \in \mathcal{D}} p(x=k|y)} \quad \text{covariance of a weighted set of instances}$$



Variational interpretation of EM

Recall: variational inference (as we saw earlier)

$$\min_q \mathbb{D}_{KL}(q(\mathbf{x})||p(\mathbf{x})) = -\mathbb{H}(q) - \mathbb{E}_q[\log p(\mathbf{x})]$$

Variational interpretation of EM

Recall: variational inference (as we saw earlier)

$$\min_q \mathbb{D}_{KL}(q(\mathbf{x})|p(\mathbf{x})) = -\mathbb{H}(q) - \mathbb{E}_q[\log p(\mathbf{x})]$$

$$p(\mathbf{x}) = \frac{\tilde{p}(\mathbf{x})}{Z}$$

Variational interpretation of EM

Recall: variational inference (as we saw earlier)

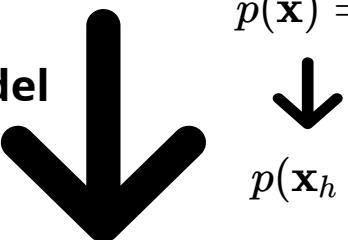
$$\min_q \mathbb{D}_{KL}(q(\mathbf{x})|p(\mathbf{x})) = -\mathbb{H}(q) - \mathbb{E}_q[\log p(\mathbf{x})] \quad \text{variational free energy}$$
$$= -\mathbb{H}(q) - \mathbb{E}_q[\log \tilde{p}(\mathbf{x})] + \log Z$$
$$p(\mathbf{x}) = \frac{\tilde{p}(\mathbf{x})}{Z}$$

Variational interpretation of EM

Recall: variational inference (as we saw earlier)

$$\min_q \mathbb{D}_{KL}(q(\mathbf{x})||p(\mathbf{x})) = -\mathbb{H}(q) - \mathbb{E}_q[\log p(\mathbf{x})] \quad \frac{\text{variational free energy}}{-\mathbb{H}(q) - \mathbb{E}_q[\log \tilde{p}(\mathbf{x})] + \log Z}$$

for a **latent variable model**

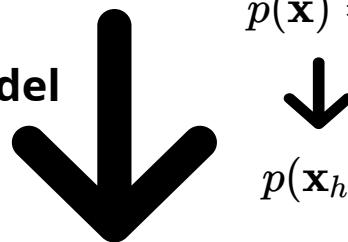

$$p(\mathbf{x}) = \frac{\tilde{p}(\mathbf{x})}{Z}$$
$$p(\mathbf{x}_h \mid \mathbf{x}_o) = \frac{p(\mathbf{x}_h, \mathbf{x}_o)}{p(\mathbf{x}_o)}$$

Variational interpretation of EM

Recall: variational inference (as we saw earlier)

$$\min_q \mathbb{D}_{KL}(q(\mathbf{x})||p(\mathbf{x})) = -\mathbb{H}(q) - \mathbb{E}_q[\log p(\mathbf{x})] \quad \frac{\text{variational free energy}}{-\mathbb{H}(q) - \mathbb{E}_q[\log \tilde{p}(\mathbf{x})] + \log Z}$$
$$p(\mathbf{x}) = \frac{\tilde{p}(\mathbf{x})}{Z}$$

for a **latent variable model**


$$p(\mathbf{x}_h \mid \mathbf{x}_o) = \frac{p(\mathbf{x}_h, \mathbf{x}_o)}{p(\mathbf{x}_o)}$$

$$\min_q \mathbb{D}_{KL}(q(\mathbf{x}_h) || p(\mathbf{x}_h \mid \mathbf{x}_o)) = -\mathbb{H}(q) - \mathbb{E}_q[\log p(\mathbf{x}_h \mid \mathbf{x}_o)]$$

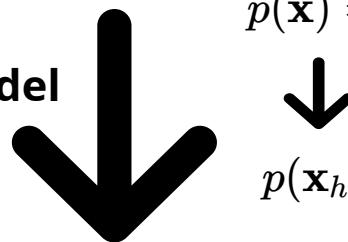
Variational interpretation of EM

Recall: variational inference (as we saw earlier)

$$\min_q \mathbb{D}_{KL}(q(\mathbf{x})||p(\mathbf{x})) = -\mathbb{H}(q) - \mathbb{E}_q[\log p(\mathbf{x})] \quad \frac{\text{variational free energy}}{-\mathbb{H}(q) - \mathbb{E}_q[\log \tilde{p}(\mathbf{x})] + \log Z}$$

$p(\mathbf{x}) = \frac{\tilde{p}(\mathbf{x})}{Z}$

for a **latent variable model**


$$p(\mathbf{x}_h \mid \mathbf{x}_o) = \frac{p(\mathbf{x}_h, \mathbf{x}_o)}{p(\mathbf{x}_o)}$$

$$\min_q \mathbb{D}_{KL}(q(\mathbf{x}_h) || p(\mathbf{x}_h \mid \mathbf{x}_o)) = -\mathbb{H}(q) - \mathbb{E}_q[\log p(\mathbf{x}_h \mid \mathbf{x}_o)]$$
$$\mathbb{E}_q[\log p(\mathbf{x}_h, \mathbf{x}_o)] - \log p(\mathbf{x}_o)$$

Variational interpretation of EM

for a **latent variable model**

$$\mathbb{D}_{KL}(q(\mathbf{x}_h) \| p(\mathbf{x}_h | \mathbf{x}_o)) = -\mathbb{H}(q) - \mathbb{E}_q[\log p(\mathbf{x}_h, \mathbf{x}_o)] + \log p(\mathbf{x}_o)$$

Variational interpretation of EM

for a **latent variable model**

$$\mathbb{D}_{KL}(q(\mathbf{x}_h) \| p(\mathbf{x}_h | \mathbf{x}_o)) = -\mathbb{H}(q) - \mathbb{E}_q[\log p(\mathbf{x}_h, \mathbf{x}_o)] + \log p(\mathbf{x}_o)$$

re-arrange 

$$\log p_\theta(\mathbf{x}_o) = \mathbb{D}_{KL}(q(\mathbf{x}_h) \| p_\theta(\mathbf{x}_h | \mathbf{x}_o)) + \mathbb{H}(q) + \mathbb{E}_q[\log p_\theta(\mathbf{x}_h, \mathbf{x}_o)]$$

original objective

non-negative, so expected log-likelihood is
a lower-bound on our original objective

expected log-likelihood wrt q

Variational interpretation of EM

for a **latent variable model**

$$\mathbb{D}_{KL}(q(\mathbf{x}_h) | p(\mathbf{x}_h | \mathbf{x}_o)) = -\mathbb{H}(q) - \mathbb{E}_q[\log p(\mathbf{x}_h, \mathbf{x}_o)] + \log p(\mathbf{x}_o)$$

re-arrange 

$$\log p_\theta(\mathbf{x}_o) = \mathbb{D}_{KL}(q(\mathbf{x}_h) | p_\theta(\mathbf{x}_h | \mathbf{x}_o)) + \mathbb{H}(q) + \mathbb{E}_q[\log p_\theta(\mathbf{x}_h, \mathbf{x}_o)]$$

original objective

non-negative, so expected log-likelihood is
a lower-bound on our original objective

expected log-likelihood wrt q

(E-step) $q(\mathbf{x}_h) = p_\theta(\mathbf{x}_h | \mathbf{x}_o)$, so the KL term is zero

(M-step) find θ to maximize the expected log-likelihood

EM as coordinate ascent

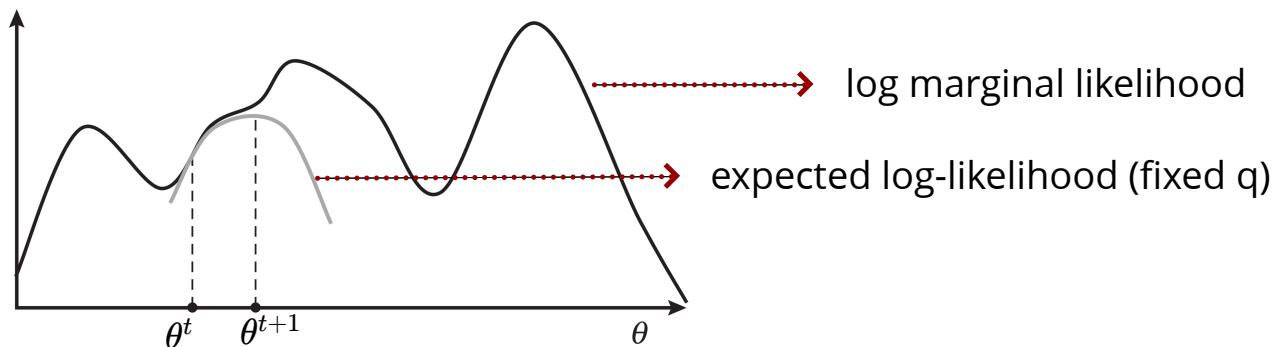
Coordinate ascent:

- E-step: optimize q for a fixed θ
- M-step: optimize θ for a fixed q

EM as coordinate ascent

Coordinate ascent:

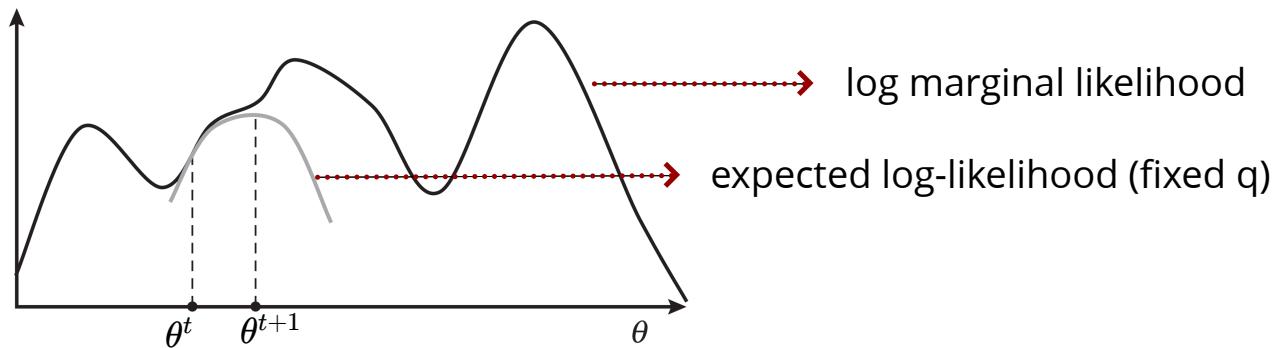
- E-step: optimize q for a fixed θ
- M-step: optimize θ for a fixed q



EM as coordinate ascent

Coordinate ascent:

- E-step: optimize q for a fixed θ
- M-step: optimize θ for a fixed q



guaranteed improvement of $\log p_\theta(\mathbf{x}_o)$ in each iteration
therefore, converges to a local optimum

Amortized inference in latent variable models

$$\log p_\theta(\mathbf{x}_o) = \mathbb{D}_{KL}(q(\mathbf{x}_h) \| p_\theta(\mathbf{x}_h | \mathbf{x}_o)) + \mathbb{H}(q) + \mathbb{E}_q[\log p_\theta(\mathbf{x}_h, \mathbf{x}_o)]$$

Amortized inference in latent variable models

$$\log p_\theta(\mathbf{x}_o) = \mathbb{D}_{KL}(q(\mathbf{x}_h)|p_\theta(\mathbf{x}_h|\mathbf{x}_o)) + \mathbb{H}(q) + \mathbb{E}_q[\log p_\theta(\mathbf{x}_h, \mathbf{x}_o)]$$

evidence lower bound (ELBO) is a lower-bound on the likelihood

Amortized inference in latent variable models

$$\log p_\theta(\mathbf{x}_o) = \mathbb{D}_{KL}(q(\mathbf{x}_h)|p_\theta(\mathbf{x}_h|\mathbf{x}_o)) + \mathbb{H}(q) + \mathbb{E}_q[\log p_\theta(\mathbf{x}_h, \mathbf{x}_o)]$$

evidence lower bound (ELBO) is a lower-bound on the likelihood



$q_\psi(\mathbf{x}_h | \mathbf{x}_o)$ instead of $q(\mathbf{x}_h)$

amortization: make q a **function** of observations

Amortized inference in latent variable models

$$\log p_\theta(\mathbf{x}_o) = \mathbb{D}_{KL}(q(\mathbf{x}_h)|p_\theta(\mathbf{x}_h|\mathbf{x}_o)) + \mathbb{H}(q) + \mathbb{E}_q[\log p_\theta(\mathbf{x}_h, \mathbf{x}_o)]$$

evidence lower bound (ELBO) is a lower-bound on the likelihood



$q_\psi(\mathbf{x}_h | \mathbf{x}_o)$ instead of $q(\mathbf{x}_h)$
amortization: make q a **function** of observations

$$\downarrow$$
$$p_\theta(\mathbf{x}_h, \mathbf{x}_o) = p_\theta(\mathbf{x}_h)p_\theta(\mathbf{x}_o | \mathbf{x}_h)$$

Amortized inference in latent variable models

$$\log p_\theta(\mathbf{x}_o) = \mathbb{D}_{KL}(q(\mathbf{x}_h)|p_\theta(\mathbf{x}_h|\mathbf{x}_o)) + \mathbb{H}(q) + \mathbb{E}_q[\log p_\theta(\mathbf{x}_h, \mathbf{x}_o)]$$

evidence lower bound (ELBO) is a lower-bound on the likelihood



$q_\psi(\mathbf{x}_h | \mathbf{x}_o)$ instead of $q(\mathbf{x}_h)$
amortization: make q a **function** of observations



$$p_\theta(\mathbf{x}_h, \mathbf{x}_o) = p_\theta(\mathbf{x}_h)p_\theta(\mathbf{x}_o | \mathbf{x}_h)$$

$$-\mathbb{D}_{KL}(q_\psi(\mathbf{x}_h | \mathbf{x}_o)|p_\theta(\mathbf{x}_h)) + \mathbb{E}_{q_\psi} [\log p_\theta(\mathbf{x}_o | \mathbf{x}_h)]$$

Amortized inference in latent variable models

$$\log p_\theta(\mathbf{x}_o) = \mathbb{D}_{KL}(q(\mathbf{x}_h)|p_\theta(\mathbf{x}_h|\mathbf{x}_o)) + \mathbb{H}(q) + \mathbb{E}_q[\log p_\theta(\mathbf{x}_h, \mathbf{x}_o)]$$

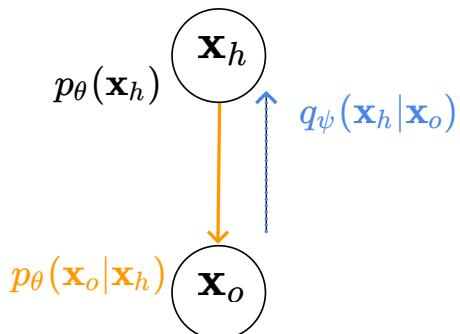
evidence lower bound (ELBO) is a lower-bound on the likelihood



$q_\psi(\mathbf{x}_h | \mathbf{x}_o)$ instead of $q(\mathbf{x}_h)$
amortization: make q a **function** of observations



$$p_\theta(\mathbf{x}_h, \mathbf{x}_o) = p_\theta(\mathbf{x}_h)p_\theta(\mathbf{x}_o | \mathbf{x}_h)$$



$$-\mathbb{D}_{KL}(q_\psi(\mathbf{x}_h | \mathbf{x}_o) | p_\theta(\mathbf{x}_h)) + \mathbb{E}_{q_\psi} [\log p_\theta(\mathbf{x}_o | \mathbf{x}_h)]$$

maximize ELBO by jointly optimizing ψ, θ

Amortized inference in latent variable models

$$\log p_\theta(\mathbf{x}_o) = \mathbb{D}_{KL}(q(\mathbf{x}_h)|p_\theta(\mathbf{x}_h|\mathbf{x}_o)) + \mathbb{H}(q) + \mathbb{E}_q[\log p_\theta(\mathbf{x}_h, \mathbf{x}_o)]$$

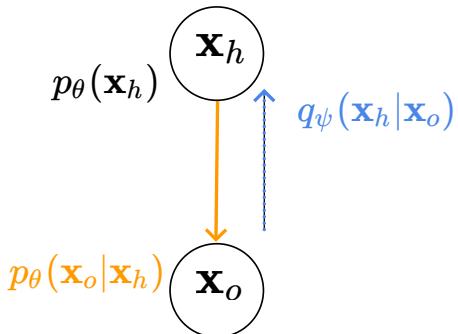
evidence lower bound (ELBO) is a lower-bound on the likelihood



$q_\psi(\mathbf{x}_h | \mathbf{x}_o)$ instead of $q(\mathbf{x}_h)$
amortization: make q a **function** of observations



$$p_\theta(\mathbf{x}_h, \mathbf{x}_o) = p_\theta(\mathbf{x}_h)p_\theta(\mathbf{x}_o|\mathbf{x}_h)$$



$$-\mathbb{D}_{KL}(q_\psi(\mathbf{x}_h | \mathbf{x}_o) | p_\theta(\mathbf{x}_h)) + \mathbb{E}_{q_\psi} [\log p_\theta(\mathbf{x}_o | \mathbf{x}_h)]$$

Variational Auto-Encoder (VAE)

maximize ELBO by jointly optimizing ψ, θ

use neural networks to represent cond. distributions

use back propagation for optimization

Undirected models with latent variables

recall

linear exponential family

$$p(x; \theta) = \frac{1}{Z(\theta)} \exp(\theta^\top \phi(x))$$

gradient in the fully observed setting

$$\nabla_\theta \ell(\theta, \mathcal{D}) = |\mathcal{D}| (\mathbb{E}_{\mathcal{D}}[\phi(x)] - \mathbb{E}_{p_\theta}[\phi(x)])$$



expectation wrt the data



expectation wrt the model

Undirected models with latent variables

recall

linear exponential family

$$p(x; \theta) = \frac{1}{Z(\theta)} \exp(\theta^\top \phi(x))$$

gradient in the fully observed setting

$$\nabla_\theta \ell(\theta, \mathcal{D}) = |\mathcal{D}| (\mathbb{E}_{\mathcal{D}}[\phi(x)] - \mathbb{E}_{p_\theta}[\phi(x)])$$



expectation wrt the data



expectation wrt the model

partial observation: $\mathbf{x} = (\mathbf{x}_o, \mathbf{x}_h)$

not observed

Undirected models with latent variables

recall

linear exponential family

$$p(x; \theta) = \frac{1}{Z(\theta)} \exp(\theta^\top \phi(x))$$

gradient in the fully observed setting

$$\nabla_\theta \ell(\theta, \mathcal{D}) = |\mathcal{D}| (\mathbb{E}_{\mathcal{D}}[\phi(x)] - \mathbb{E}_{p_\theta}[\phi(x)])$$



expectation wrt the data



expectation wrt the model

partial observation: $\mathbf{x} = (\mathbf{x}_o, \mathbf{x}_h)$

not observed

marginal likelihood: $p(\mathbf{x}_o; \theta) = \sum_{\mathbf{x}_h} \frac{1}{Z(\theta)} \exp(\theta^\top \phi(\mathbf{x}))$

Undirected models with latent variables

recall

linear exponential family

$$p(x; \theta) = \frac{1}{Z(\theta)} \exp(\theta^\top \phi(x))$$

gradient in the fully observed setting

$$\nabla_\theta \ell(\theta, \mathcal{D}) = |\mathcal{D}| (\mathbb{E}_{\mathcal{D}}[\phi(x)] - \mathbb{E}_{p_\theta}[\phi(x)])$$



expectation wrt the data

expectation wrt the model

partial observation: $\mathbf{x} = (\mathbf{x}_o, \mathbf{x}_h)$

not observed

marginal likelihood: $p(\mathbf{x}_o; \theta) = \sum_{\mathbf{x}_h} \frac{1}{Z(\theta)} \exp(\theta^\top \phi(\mathbf{x}))$

gradient in the partially obs. case

$$\nabla_\theta \ell(\theta, \mathcal{D}) = |\mathcal{D}| (\mathbb{E}_{\mathcal{D}, \theta}[\phi(x)] - \mathbb{E}_{p_\theta}[\phi(x)])$$

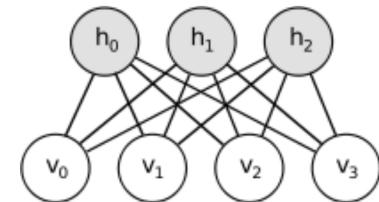


wrt both data and model: we need to do inference to calculate expected sufficient statistics (similar to E-step in EM)

Example: Restricted Boltzmann Machine (RBM)

binary RBM: $p(h, v) = \frac{1}{Z(\theta)} \exp(\sum_{i,j} \theta_{i,j} v_i h_j)$

data: $\mathcal{D} = \{v^{(m)}\}_m$ for $v_i, h_j \in \{0, 1\}$

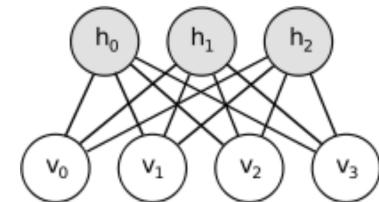


Example: Restricted Boltzmann Machine (RBM)

binary RBM: $p(h, v) = \frac{1}{Z(\theta)} \exp(\sum_{i,j} \theta_{i,j} v_i h_j)$

data: $\mathcal{D} = \{v^{(m)}\}_m$ for $v_i, h_j \in \{0, 1\}$

sufficient statistics: $\phi(v_i, h_j) = v_i h_j$



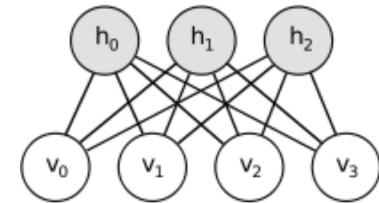
Example: Restricted Boltzmann Machine (RBM)

binary RBM: $p(h, v) = \frac{1}{Z(\theta)} \exp(\sum_{i,j} \theta_{i,j} v_i h_j)$

data: $\mathcal{D} = \{v^{(m)}\}_m$ for $v_i, h_j \in \{0, 1\}$

sufficient statistics: $\phi(v_i, h_j) = v_i h_j$

we want to optimize: $\ell(\mathcal{D}; \theta) = \sum_{v \in \mathcal{D}} \log \sum_h \frac{1}{Z(\theta)} \exp(\sum_{i,j} \theta_{i,j} v_i h_j)$



Example: Restricted Boltzmann Machine (RBM)

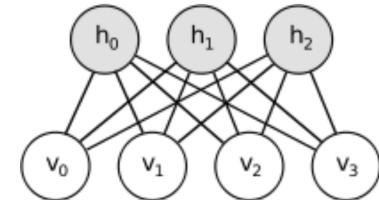
binary RBM: $p(h, v) = \frac{1}{Z(\theta)} \exp(\sum_{i,j} \theta_{i,j} v_i h_j)$

data: $\mathcal{D} = \{v^{(m)}\}_m$ for $v_i, h_j \in \{0, 1\}$

sufficient statistics: $\phi(v_i, h_j) = v_i h_j$

we want to optimize: $\ell(\mathcal{D}; \theta) = \sum_{v \in \mathcal{D}} \log \sum_h \frac{1}{Z(\theta)} \exp(\sum_{i,j} \theta_{i,j} v_i h_j)$

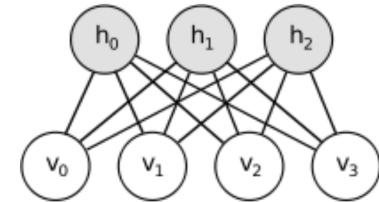
gradient: $\frac{\partial}{\partial \theta_{i,j}} \ell(\mathcal{D}; \theta) \propto \mathbb{E}_{\mathcal{D}, \theta}[v_i h_j] - \mathbb{E}_{p_\theta}[v_i h_j]$
 $= (\frac{1}{M} \sum_{v'_i \in \mathcal{D}} v'_i \mathbb{E}_{p_\theta}[h_j | v'_i]) - \mathbb{E}_{p_\theta}[v_i h_j]$



Example: Restricted Boltzmann Machine (RBM)

binary RBM: $p(h, v) = \frac{1}{Z(\theta)} \exp(\sum_{i,j} \theta_{i,j} v_i h_j)$

data: $\mathcal{D} = \{v^{(m)}\}_m$ for $v_i, h_j \in \{0, 1\}$



sufficient statistics: $\phi(v_i, h_j) = v_i h_j$

we want to optimize: $\ell(\mathcal{D}; \theta) = \sum_{v \in \mathcal{D}} \log \sum_h \frac{1}{Z(\theta)} \exp(\sum_{i,j} \theta_{i,j} v_i h_j)$

gradient: $\frac{\partial}{\partial \theta_{i,j}} \ell(\mathcal{D}; \theta) \propto \mathbb{E}_{\mathcal{D}, \theta}[v_i h_j] - \mathbb{E}_{p_\theta}[v_i h_j]$

$$= (\frac{1}{M} \sum_{v'_i \in \mathcal{D}} v'_i \mathbb{E}_{p_\theta}[h_j | v'_i]) - \mathbb{E}_{p_\theta}[v_i h_j]$$

sampling-based inference: sample $h \mid v$

use Gibbs sampling:
sample both h, v using current parameters

Summary

learning with partial observations:

- missing data
 - optimize the log-marginal-likelihood when **missing at random**
- latent variables
 - can produce expressive probabilistic models

problem is not convex

how to learn the model?

- directly estimate the gradient (*directed and undirected*)
- use EM (*directed models*)
 - variational interpretation + relation to ELBO and VAEs