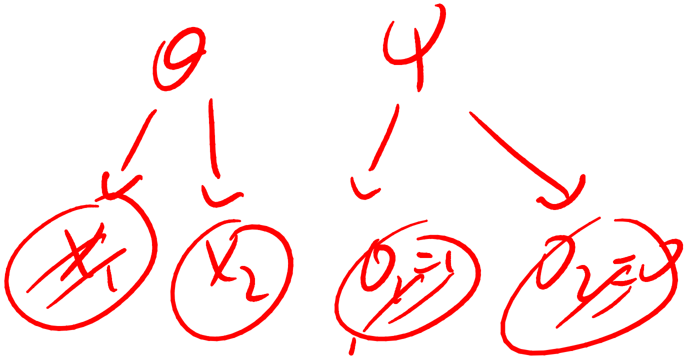
Stat 521A
Lecture 23

Outline

- Basic issues (19.1)
- Gradient ascent for DGMs (19.2.1)
- EM for DGMs (19.2.2)
- Variational EM (19.2.4)
- MCMC for param inf in DGMs (19.3.2)
- Variational Bayes (19.3.3)

MCAR

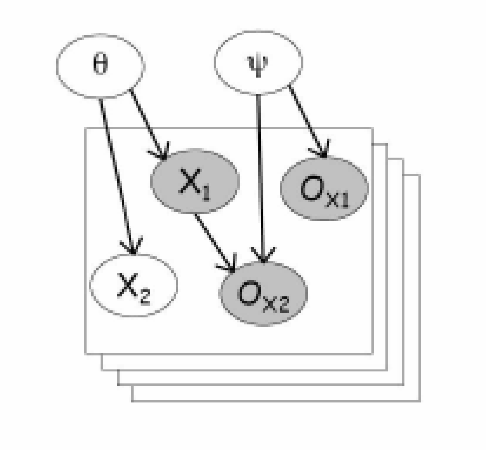
- Let Xi be the true value of variable I, and Oi in {0,1} be whether it is observed or not. Yi(Oi) = Xi or ?.
- Defn 19.1.6. Missing completely at random (MCAR) means X \perp O.
- Given MCAR, we can safely ignore the missing variables (for which Ox=1), since they tell us nothing about theta



$$p(\theta, \psi | Y_1, Y_2) = p(\theta | X_1) p(\psi | O_1, O_2)$$

Missing at random

- Defn 19.1.8. Let H be hidden vars, V be visible vars, and O be observation status. Missing at random means O \perp H | V.



- Intuitively, although O may depend on some of the variables Xv, since we observe Xv, we do not learn anything new about Xh.

Benefits of MAR

- Thm 19.1.9. Given MAR, and a factored prior, p(θ,ψ|D) = p(θ|Xv) p(ψ|Xv,O)
- Pf.

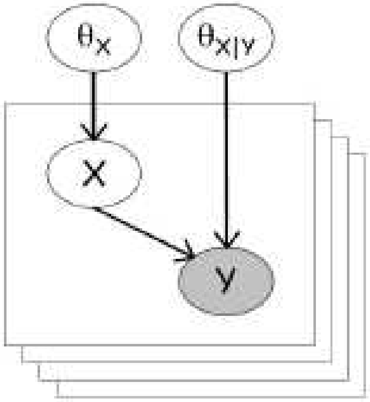
$$
\begin{aligned}
P_{missing}(\boldsymbol{y}) &= \sum_{\boldsymbol{x}^{\boldsymbol{y}}_{\text{hidden}}} \left[P(\boldsymbol{x}^{\boldsymbol{y}}_{\text{obs}}, \boldsymbol{x}^{\boldsymbol{y}}_{\text{hidden}}) P_{missing}(o_{\mathbf{X}} \mid \boldsymbol{x}^{\boldsymbol{y}}_{\text{hidden}}, \boldsymbol{x}^{\boldsymbol{y}}_{\text{obs}})\right] \\
&= \sum_{\boldsymbol{x}^{\boldsymbol{y}}_{\text{hidden}}} \left[P(\boldsymbol{x}^{\boldsymbol{y}}_{\text{obs}}, \boldsymbol{x}^{\boldsymbol{y}}_{\text{hidden}}) P_{missing}(o_{\mathbf{X}} \mid \boldsymbol{x}^{\boldsymbol{y}}_{\text{obs}})\right] \\
&= P_{missing}(o_{\mathbf{X}} \mid \boldsymbol{x}^{\boldsymbol{y}}_{\text{obs}}) \sum_{\boldsymbol{x}^{\boldsymbol{y}}_{\text{hidden}}} P(\boldsymbol{x}^{\boldsymbol{y}}_{\text{obs}}, \boldsymbol{x}^{\boldsymbol{y}}_{\text{hidden}}) \\
&= \underset{\overset{\uparrow}{\psi}}{P_{missing}}(o_{\mathbf{X}} \mid \boldsymbol{x}^{\boldsymbol{y}}_{\text{obs}}) \underset{\overset{\uparrow}{\theta}}{P}(\boldsymbol{x}^{\boldsymbol{y}}_{\text{obs}}).
\end{aligned}
$$

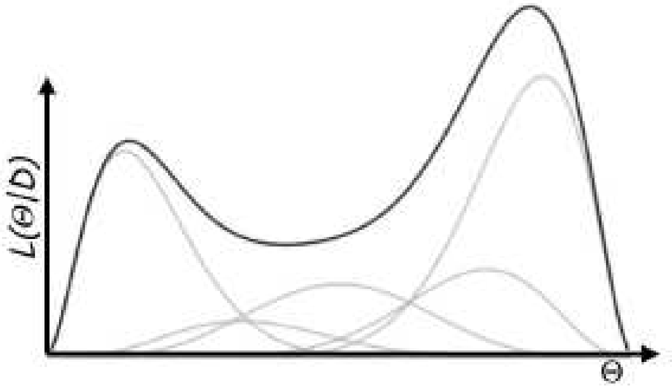
Counter examples to MAR

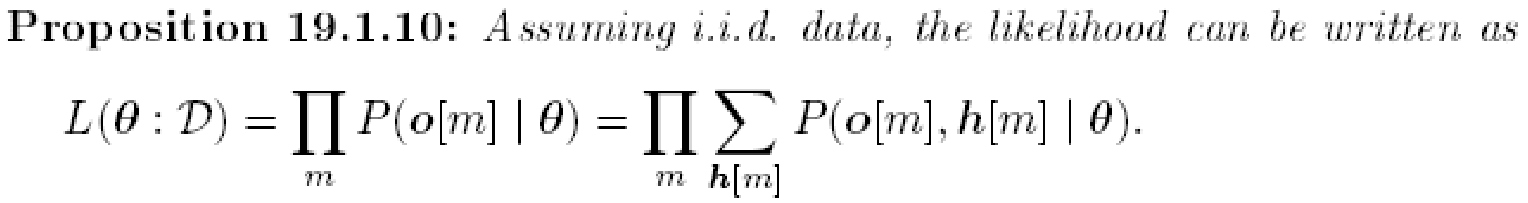
- Collaborative filtering: people are more likely to rate movies they strongly like or dislike.
- Medicine: if a patient does not have a check mark in the “had X-ray” field, they probably don’t have any bone problems. However, if we explicitly write the “primary complaint” as the cause of which tests are performed, MAR is restored (since we observe why O(Xray)=0).
- Henceforth we will assume MAR.

Multimodality

- For fully observed DGMs, likelihood is convex (assuming each CPD is convex), and hence has a single global maximum.
- When we have missing data, the likelihood is a mixture of up to K^n modes, corresponding to every possible completion pattern





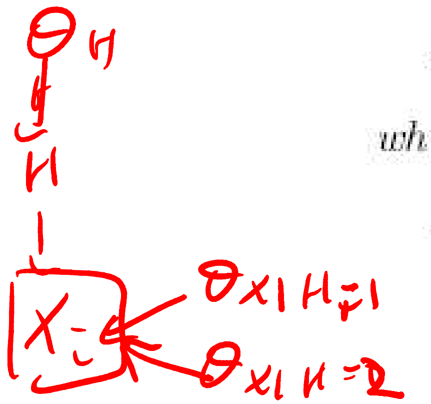


Proposition 19.1.10: *Assuming i.i.d. data, the likelihood can be written as*

$$L(\boldsymbol{\theta} : \mathcal{D}) = \prod_m P(\boldsymbol{o}[m] \mid \boldsymbol{\theta}) = \prod_m \sum_{\boldsymbol{h}[m]} P(\boldsymbol{o}[m], \boldsymbol{h}[m] \mid \boldsymbol{\theta}).$$

Identifiability

- Sometimes we cannot uniquely identify the parameters, even given infinite data
- Eg The experimenter either tosses coin 1 or coin 2, but we don't know which. The model is



$$L(\theta : \mathcal{D}) = P(X = Heads)^{M[Heads]}(1 - P(X = Heads))^{M[Tails]},$$

where

$$P(X = Heads) = \theta_H \theta_{X|h^1} + (1 - \theta_H)\theta_{X|h^2}.$$

- We have eg. $p(D|\theta_H=0.5, \theta_1=0.5, \theta_2=0.5) = p(D|\theta_H=0.5, \theta_1=0.8, \theta_2=0.2)$. The problem is underconstrained.

Identifiability

- Defn 19.1.13. A parameter θ is identifiable if there is no $\theta' \neq \theta$ st $p(X|\theta)=p(X|\theta')$. A model is identifiable if all θ are identifiable.
- A mixture model cannot be identifiable since we can always arbitrarily permute the hidden labels, and the corresponding parameters.
- Hence we should not ask things like “what is the prob. Xi belongs to cluster k” but rather “what is the prob Xi and Xi belong to the same cluster”.

Gradient descent for DGMs

- We can find a local maximum using gradient based methods.
- Consider tabular CPDs.
- Thm 19.2.1.

$$\frac{\partial}{\partial \theta_{ijk}} p(\mathbf{e}) = \frac{p(x_i = k, \mathbf{x}_{\pi_i} = j, \mathbf{e})}{\theta_{ijk}}$$

- Pf.

$$\begin{aligned} \frac{\partial}{\partial \theta_{ijk}} \prod_{i'} \theta_{i', \mathbf{x}_{i'}, x_{i'}} &= \prod_{i' \neq i} \theta_{i', \mathbf{x}_{i'}, x_{i'}} I(\mathbf{x}_i = j, x_i = k) \\ &= \frac{p(\mathbf{e})}{\theta_{ijk}} I(\mathbf{x}_i = j, x_i = k) \end{aligned}$$

Gradient descent for DGMs

- Pf contd

$$\begin{aligned}
\frac{\partial}{\partial P(x \mid \boldsymbol{u})} P(e) &= \sum_{\xi : \xi\langle \boldsymbol{E} \rangle = \boldsymbol{e}} \frac{\partial}{\partial P(x \mid \boldsymbol{u})} P(\xi) \\
&= \sum_{\xi : \xi\langle \boldsymbol{E} \rangle = \boldsymbol{e}, \xi\langle X, \mathrm{Pa}_X \rangle = \langle x, \boldsymbol{u} \rangle} \frac{1}{P(x \mid \boldsymbol{u})} P(\xi) \\
&= \frac{1}{P(x \mid \boldsymbol{u})} P(x, \boldsymbol{u}, \boldsymbol{e}).
\end{aligned}$$

- Thm 19.2.2.

$$\frac{\partial \ell(\boldsymbol{\theta} : \mathcal{D})}{\partial P(x \mid \boldsymbol{u})} = \frac{1}{P(x \mid \boldsymbol{u})} \sum_{m=1}^{M} P(x, \boldsymbol{u} \mid \boldsymbol{o}[m], \boldsymbol{\theta}).$$

- Chain rule for non-tabular CPDs.

$$\frac{\partial \ell(\boldsymbol{\theta} : \mathcal{D})}{\partial \theta} = \sum_{x, \boldsymbol{u}} \frac{\partial \ell(P_{\boldsymbol{\theta}} : \mathcal{D})}{\partial P(x \mid \boldsymbol{u})} \frac{\partial P(x \mid \boldsymbol{u})}{\partial \theta},$$

Gradient algorithms

- Gradient requires inference to compute family marginals.
- Need to enforce positivity and sum-to-one constraints (for discrete) eg reparameterize to unconstrained form

$$P(x \mid \boldsymbol{u}) = \frac{e^{\lambda_{x|\boldsymbol{u}}}}{\sum_{x' \in Val(X)} e^{\lambda_{x'|\boldsymbol{u}}}}.$$

- Need to enforce positive definite – optimize wrt the cholesky factors.
- Have to specify step-size and search direction (use black-box algorithm).
- EM is much easier…

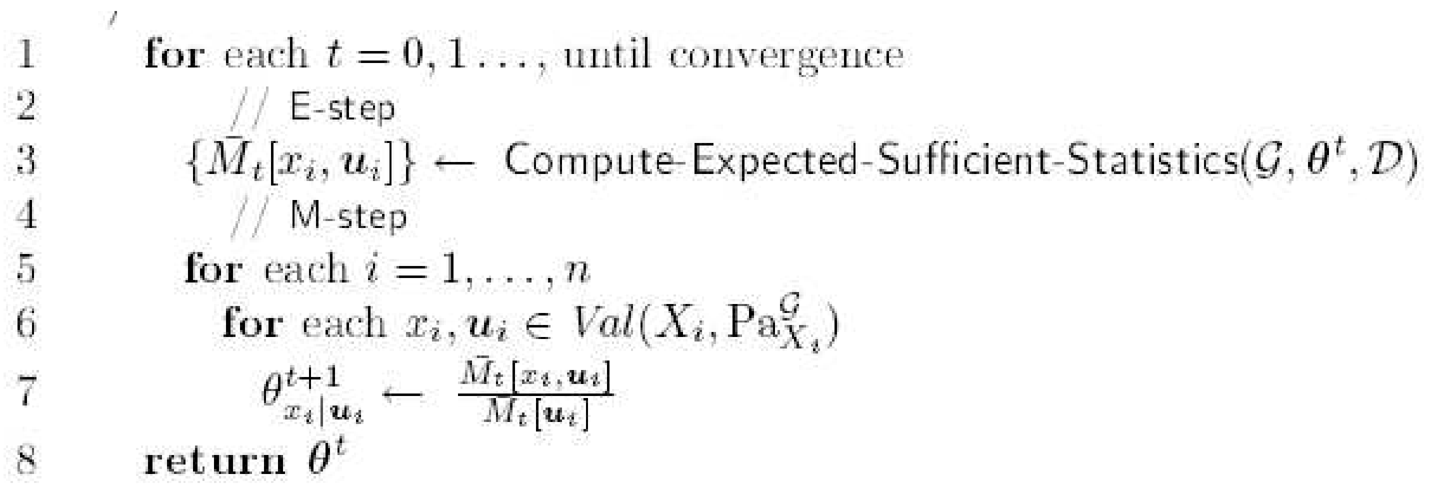
EM for DGMs

- Key intuition: if we knew the values of H, we could compute the MLEs/MAP estimates for θ easily. So we infer H|θ and then estimate θ|H. For the latter, we just need the expected sufficient statistics. For tabular CPDs, this is just a table of expcted counts
- E step

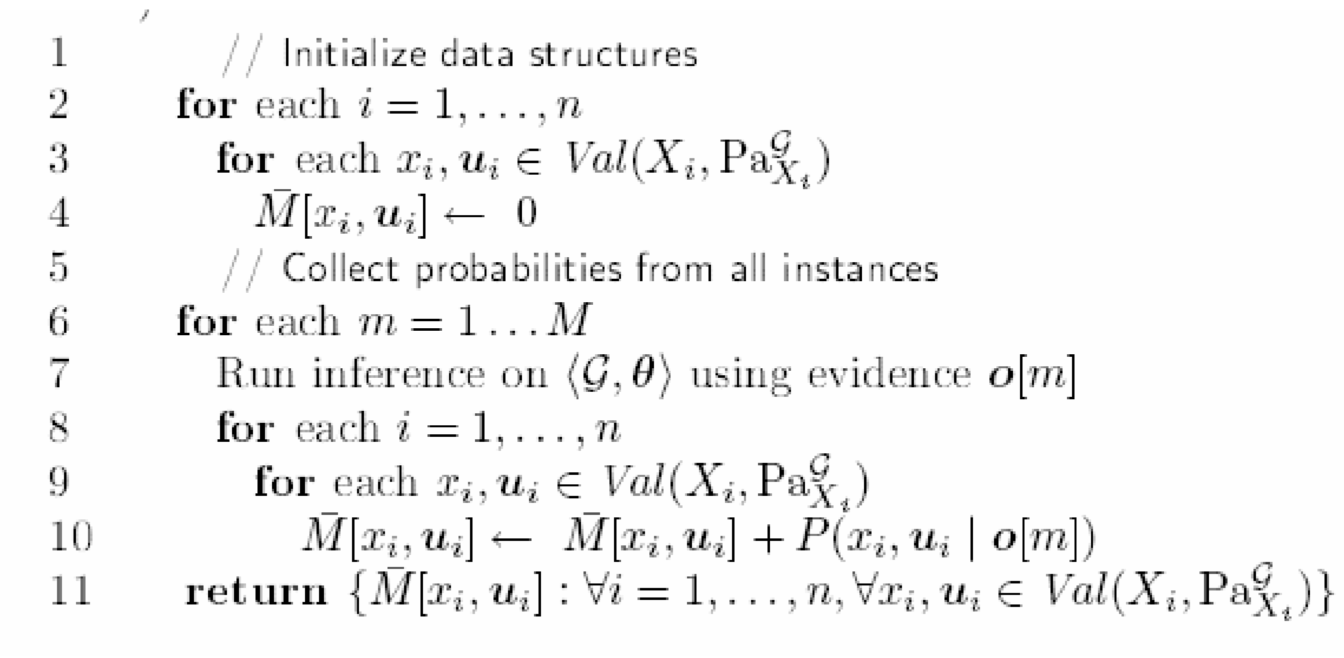
$$\bar{M}_{\boldsymbol{\theta}^t}[x, \boldsymbol{u}] = \sum_m P(x, \boldsymbol{u} \mid \boldsymbol{o}[m], \boldsymbol{\theta}^t).$$

- M step
- $$\theta^{t+1}_{x|\boldsymbol{u}} = \frac{\bar{M}_{\boldsymbol{\theta}^t}[x, \boldsymbol{u}]}{\bar{M}_{\boldsymbol{\theta}^t}[\boldsymbol{u}]}$$

Pseudocode



```
for each $t = 0, 1 \ldots,$ until convergence
  // E-step
  $\{\bar{M}_t[x_i, \boldsymbol{u}_i]\} \leftarrow$ Compute-Expected-Sufficient-Statistics($\mathcal{G}, \boldsymbol{\theta}^t, \mathcal{D}$)
  // M-step
  for each $i = 1, \ldots, n$
    for each $x_i, \boldsymbol{u}_i \in Val(X_i, \mathrm{Pa}^{\mathcal{G}}_{X_i})$
      $\theta^{t+1}_{x_i|\boldsymbol{u}_i} \leftarrow \frac{\bar{M}_t[x_i, \boldsymbol{u}_i]}{\bar{M}_t[\boldsymbol{u}_i]}$
return $\theta^t$
```



```
  // Initialize data structures
  for each $i = 1, \ldots, n$
    for each $x_i, \boldsymbol{u}_i \in Val(X_i, \mathrm{Pa}^{\mathcal{G}}_{X_i})$
      $\bar{M}[x_i, \boldsymbol{u}_i] \leftarrow 0$
  // Collect probabilities from all instances
  for each $m = 1 \ldots M$
    Run inference on $\langle \mathcal{G}, \boldsymbol{\theta} \rangle$ using evidence $\boldsymbol{o}[m]$
    for each $i = 1, \ldots, n$
      for each $x_i, \boldsymbol{u}_i \in Val(X_i, \mathrm{Pa}^{\mathcal{G}}_{X_i})$
        $\bar{M}[x_i, \boldsymbol{u}_i] \leftarrow \bar{M}[x_i, \boldsymbol{u}_i] + P(x_i, \boldsymbol{u}_i \mid \boldsymbol{o}[m])$
  return $\{\bar{M}[x_i, \boldsymbol{u}_i] : \forall i = 1, \ldots, n, \forall x_i, \boldsymbol{u}_i \in Val(X_i, \mathrm{Pa}^{\mathcal{G}}_{X_i})\}$
```

ECDLL

- Define expected complete data log likelihood, wrt Q distribution over H|D

$$\boldsymbol{E}_Q[\ell(\boldsymbol{\theta} : \langle \mathcal{D}, \mathcal{H} \rangle)] = \sum_{\mathcal{H}} Q(\mathcal{H}) \ell(\boldsymbol{\theta} : \langle \mathcal{D}, \mathcal{H} \rangle)$$

- For tabular CPDs, we have

$$\ell(\boldsymbol{\theta} : \langle \mathcal{D}, \mathcal{H} \rangle) = \sum_{i=1}^{n} \sum_{(x_i, \boldsymbol{u}_i) \in Val(X_i, \mathrm{Pa}_{X_i})} M_{\langle \mathcal{D}, \mathcal{H} \rangle}[x_i, \boldsymbol{u}_i] \log \theta_{x_i | \boldsymbol{u}_i}.$$

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

$$\boldsymbol{E}_Q[\ell(\boldsymbol{\theta} : \langle \mathcal{D}, \mathcal{H} \rangle)] = \sum_{i=1}^{n} \sum_{(x_i, \boldsymbol{u}_i) \in Val(X_i, \mathrm{Pa}_{X_i})} \bar{M}_Q[x_i, \boldsymbol{u}_i] \log \theta_{x_i | \boldsymbol{u}_i}.$$

ECDLL for exp fam

- The key to making EM simple for expfam models is that the log-likelihood is linear in the sufficient statistics

$$P(\xi \mid \theta) = \frac{1}{Z(\theta)} A(\xi) \exp\left\{ \langle \mathrm{t}(\theta), \tau(\xi) \rangle \right\}, \qquad \tau(\langle \mathcal{D}, \mathcal{H} \rangle) = \sum_m \tau(o[m], h[m]).$$

$$\ell(\theta : \langle \mathcal{D}, \mathcal{H} \rangle) = \langle \mathrm{t}(\theta), \tau(\langle \mathcal{D}, \mathcal{H} \rangle) \rangle + \sum_m A(o[m], h[m]) - \log Z(\theta).$$

$$\boldsymbol{E}_Q[\ell(\theta : \langle \mathcal{D}, \mathcal{H} \rangle)] = \langle \mathrm{t}(\theta), \boldsymbol{E}_Q[\tau(\langle \mathcal{D}, \mathcal{H} \rangle)] \rangle + \sum_m \boldsymbol{E}_Q[\underbrace{\cancel{A(o[m], h[m])}}_{\text{const}}] - M \log Z(\theta).$$

Choosing Q (for E step)

- Define

$$\Phi_{\mathcal{D}}[\theta, Q] = \mathbb{E}_Q[\ell(\theta : \langle \mathcal{D}, \mathcal{H} \rangle)] + \mathbb{H}_Q(\mathcal{H}).$$

- Thm 19.2.5.

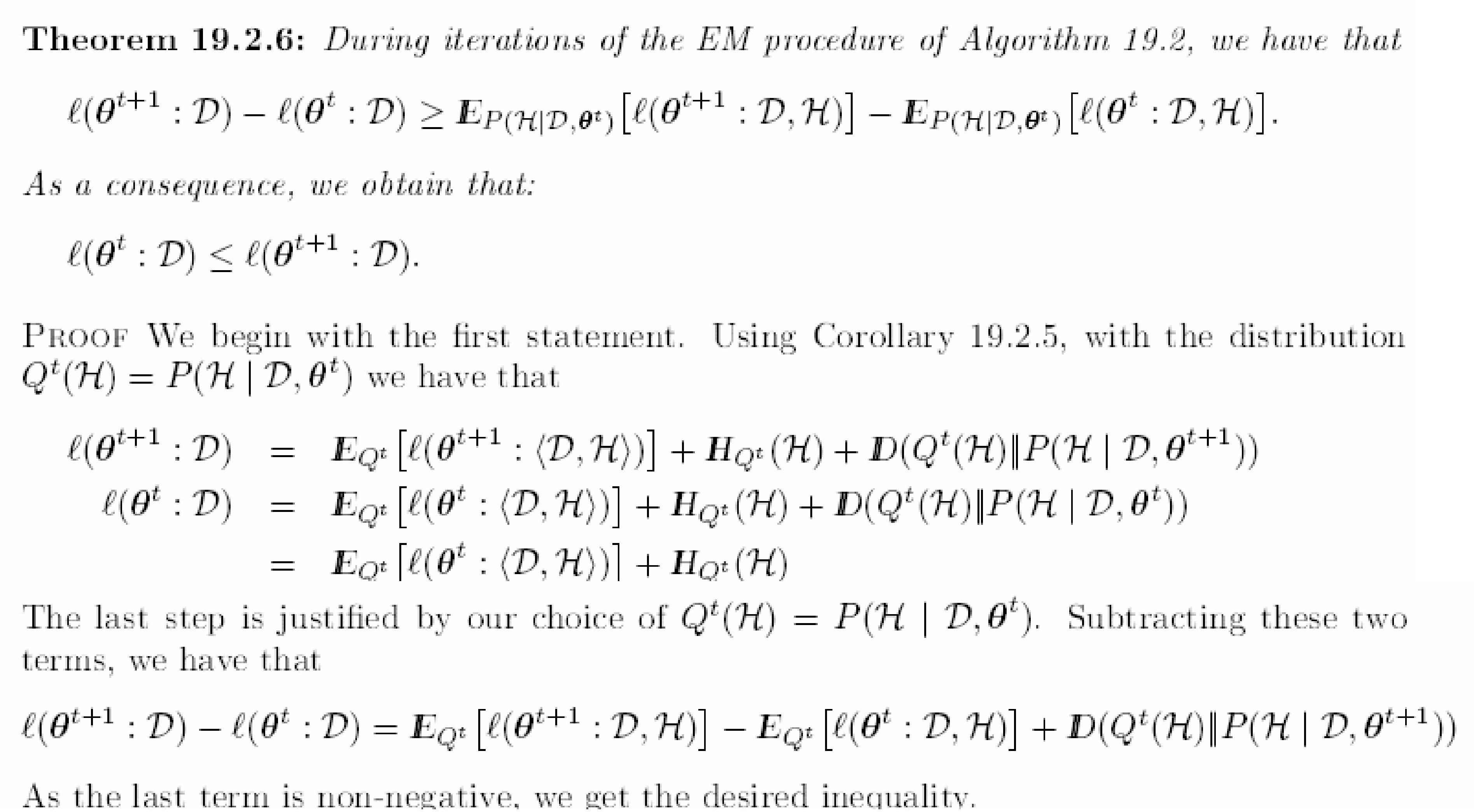
Corollary 19.2.5: *For any* Q,

$$\begin{aligned} \ell(\theta : \mathcal{D}) &= \Phi_{\mathcal{D}}[\theta, Q] + \mathbb{D}(Q(\mathcal{H}) \| P(\mathcal{H} \mid \mathcal{D}, \theta)) \\ &= \mathbb{E}_Q[\ell(\theta : \langle \mathcal{D}, \mathcal{H} \rangle)] + \mathbb{H}_Q(\mathcal{H}) + \mathbb{D}(Q(\mathcal{H}) \| P(\mathcal{H} \mid \mathcal{D}, \theta)). \end{aligned}$$

- From (2), ECDLL is lower bound on LL.
- From (1), if Q=p(H|D,\theta), then bound is tight.
- EM alternates between optimizing Q and optimizing \theta. Can do partial updates.

Convergence

- Thm 19.2.6. If we do exact EM (so Q=p(H|D,theta)), then the LL never decreases



Theorem 19.2.6: *During iterations of the EM procedure of Algorithm 19.2, we have that*

$$\ell(\boldsymbol{\theta}^{t+1} : \mathcal{D}) - \ell(\boldsymbol{\theta}^{t} : \mathcal{D}) \geq \boldsymbol{E}_{P(\mathcal{H}|\mathcal{D},\boldsymbol{\theta}^t)}\left[\ell(\boldsymbol{\theta}^{t+1} : \mathcal{D},\mathcal{H})\right] - \boldsymbol{E}_{P(\mathcal{H}|\mathcal{D},\boldsymbol{\theta}^t)}\left[\ell(\boldsymbol{\theta}^{t} : \mathcal{D},\mathcal{H})\right].$$

As a consequence, we obtain that:

$$\ell(\boldsymbol{\theta}^{t} : \mathcal{D}) \leq \ell(\boldsymbol{\theta}^{t+1} : \mathcal{D}).$$

Proof We begin with the first statement. Using Corollary 19.2.5, with the distribution $Q^t(\mathcal{H}) = P(\mathcal{H} \mid \mathcal{D}, \boldsymbol{\theta}^t)$ we have that

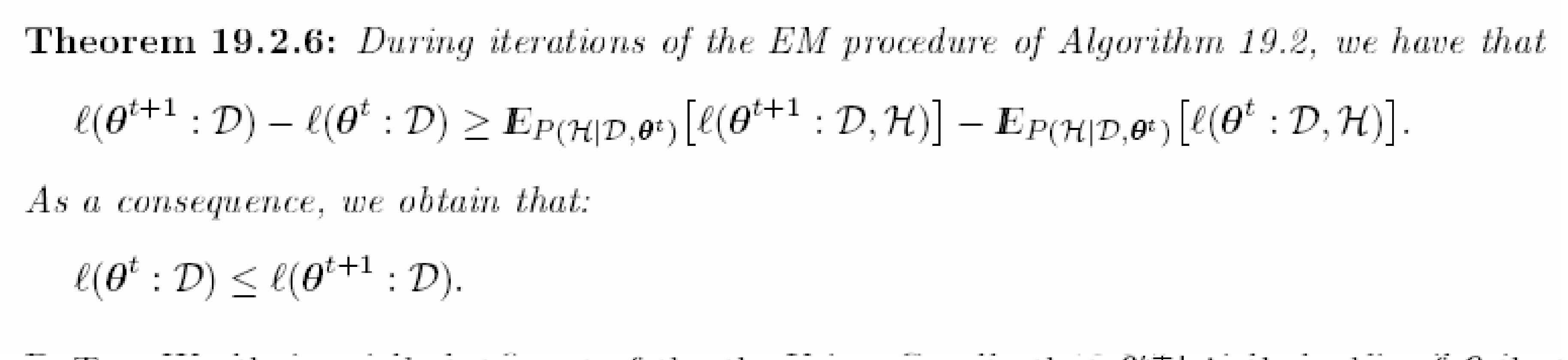
$$\begin{aligned}
\ell(\boldsymbol{\theta}^{t+1} : \mathcal{D}) &= \boldsymbol{E}_{Q^t}\left[\ell(\boldsymbol{\theta}^{t+1} : \langle\mathcal{D},\mathcal{H}\rangle)\right] + \boldsymbol{H}_{Q^t}(\mathcal{H}) + \boldsymbol{D}(Q^t(\mathcal{H})\|P(\mathcal{H} \mid \mathcal{D},\boldsymbol{\theta}^{t+1})) \\
\ell(\boldsymbol{\theta}^{t} : \mathcal{D}) &= \boldsymbol{E}_{Q^t}\left[\ell(\boldsymbol{\theta}^{t} : \langle\mathcal{D},\mathcal{H}\rangle)\right] + \boldsymbol{H}_{Q^t}(\mathcal{H}) + \boldsymbol{D}(Q^t(\mathcal{H})\|P(\mathcal{H} \mid \mathcal{D},\boldsymbol{\theta}^{t})) \\
&= \boldsymbol{E}_{Q^t}\left[\ell(\boldsymbol{\theta}^{t} : \langle\mathcal{D},\mathcal{H}\rangle)\right] + \boldsymbol{H}_{Q^t}(\mathcal{H})
\end{aligned}$$

The last step is justified by our choice of $Q^t(\mathcal{H}) = P(\mathcal{H} \mid \mathcal{D}, \boldsymbol{\theta}^t)$. Subtracting these two terms, we have that

$$\ell(\boldsymbol{\theta}^{t+1} : \mathcal{D}) - \ell(\boldsymbol{\theta}^{t} : \mathcal{D}) = \boldsymbol{E}_{Q^t}\left[\ell(\boldsymbol{\theta}^{t+1} : \mathcal{D},\mathcal{H})\right] - \boldsymbol{E}_{Q^t}\left[\ell(\boldsymbol{\theta}^{t} : \mathcal{D},\mathcal{H})\right] + \boldsymbol{D}(Q^t(\mathcal{H})\|P(\mathcal{H} \mid \mathcal{D},\boldsymbol{\theta}^{t+1}))$$

As the last term is non-negative, we get the desired inequality.

Convergence cont'd

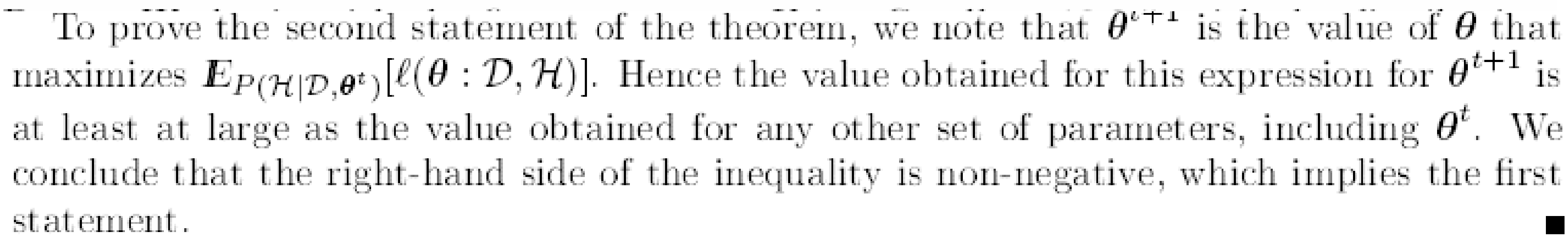


Theorem 19.2.6: *During iterations of the EM procedure of Algorithm 19.2, we have that*

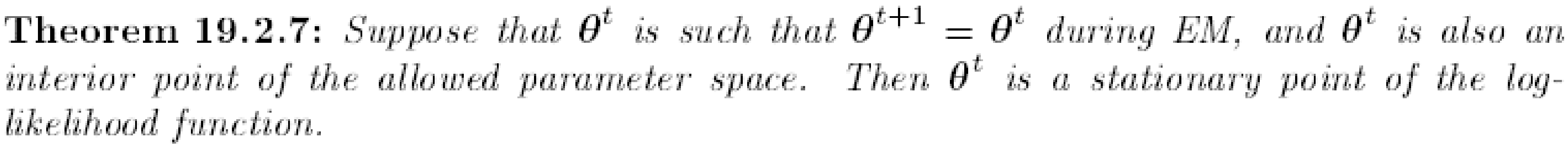
$$\ell(\boldsymbol{\theta}^{t+1} : \mathcal{D}) - \ell(\boldsymbol{\theta}^{t} : \mathcal{D}) \geq \boldsymbol{E}_{P(\mathcal{H}|\mathcal{D},\boldsymbol{\theta}^t)}\left[\ell(\boldsymbol{\theta}^{t+1} : \mathcal{D}, \mathcal{H})\right] - \boldsymbol{E}_{P(\mathcal{H}|\mathcal{D},\boldsymbol{\theta}^t)}\left[\ell(\boldsymbol{\theta}^{t} : \mathcal{D}, \mathcal{H})\right].$$

As a consequence, we obtain that:

$$\ell(\boldsymbol{\theta}^{t} : \mathcal{D}) \leq \ell(\boldsymbol{\theta}^{t+1} : \mathcal{D}).$$



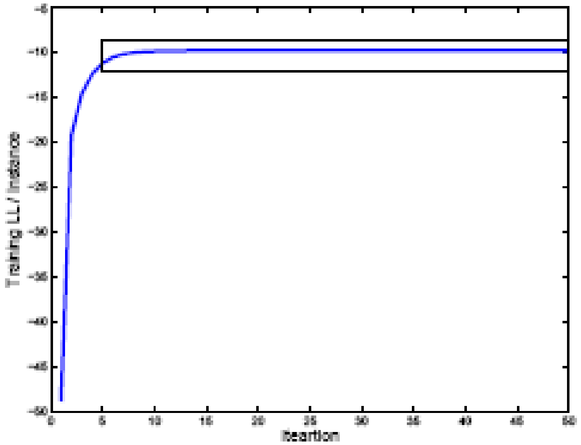
To prove the second statement of the theorem, we note that $\boldsymbol{\theta}^{t+1}$ is the value of $\boldsymbol{\theta}$ that maximizes $\boldsymbol{E}_{P(\mathcal{H}|\mathcal{D},\boldsymbol{\theta}^t)}[\ell(\boldsymbol{\theta} : \mathcal{D}, \mathcal{H})]$. Hence the value obtained for this expression for $\boldsymbol{\theta}^{t+1}$ is at least at large as the value obtained for any other set of parameters, including $\boldsymbol{\theta}^{t}$. We conclude that the right-hand side of the inequality is non-negative, which implies the first statement. ■



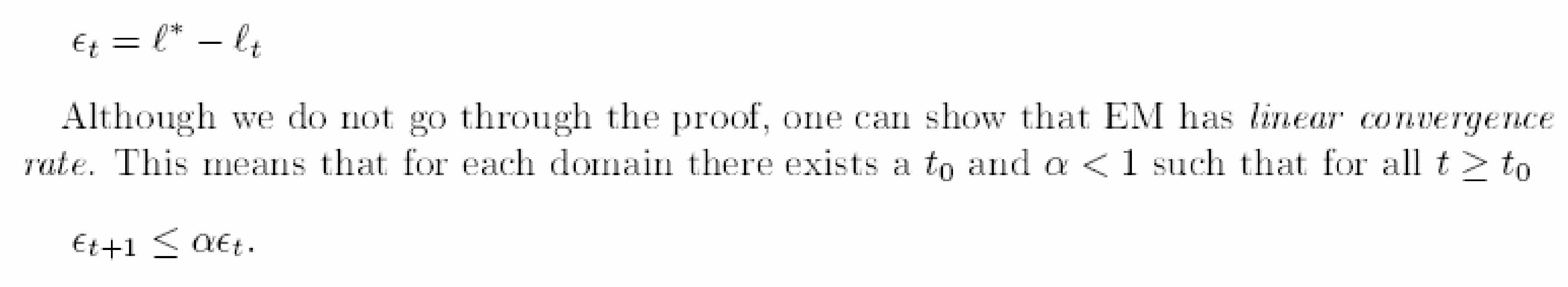
Theorem 19.2.7: *Suppose that $\boldsymbol{\theta}^{t}$ is such that $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^{t}$ during EM, and $\boldsymbol{\theta}^{t}$ is also an interior point of the allowed parameter space. Then $\boldsymbol{\theta}^{t}$ is a stationary point of the log-likelihood function.*

Rate of convergence

- Initially fast, then very slow; can switch over to conjugate gradient near optimum



- EM has linear convergence rate



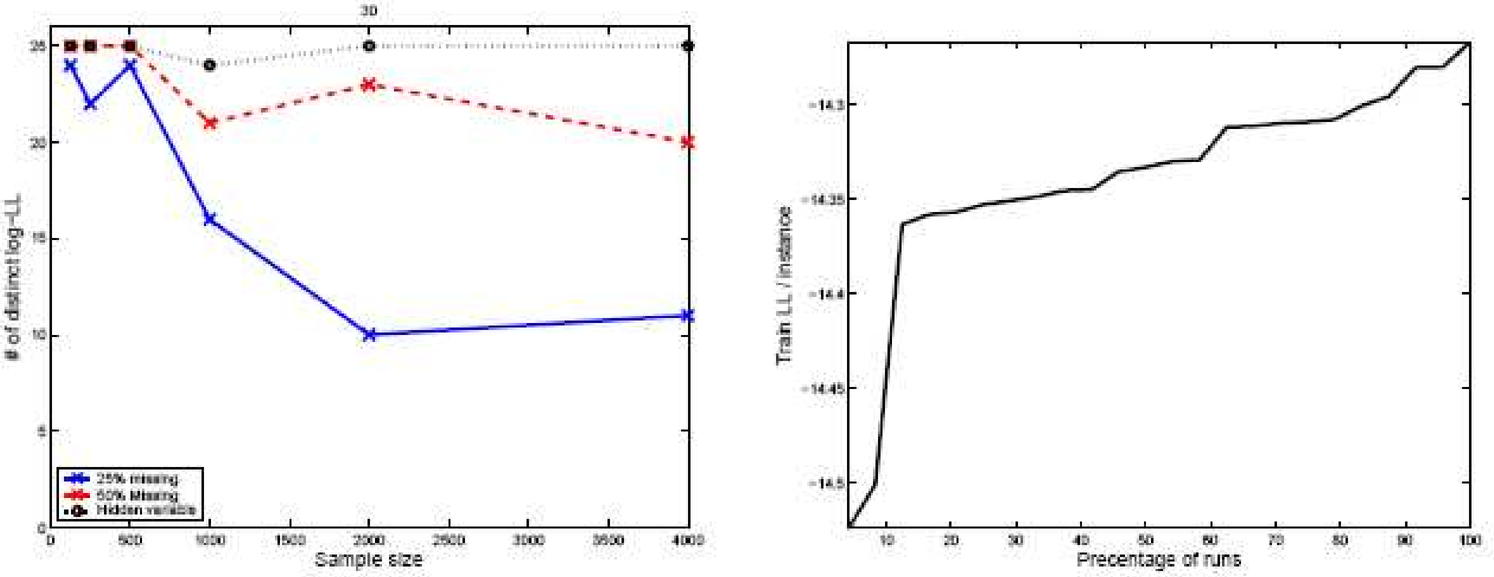
$\epsilon_t = \ell^* - \ell_t$

Although we do not go through the proof, one can show that EM has *linear convergence rate*. This means that for each domain there exists a t_0 and $\alpha < 1$ such that for all $t \geq t_0$

$\epsilon_{t+1} \leq \alpha \epsilon_t$.

Local maxima

- Maxima can differ a lot in quality.
- Can do multiple restart, killing off some runs early if they look bad (as in beam search).



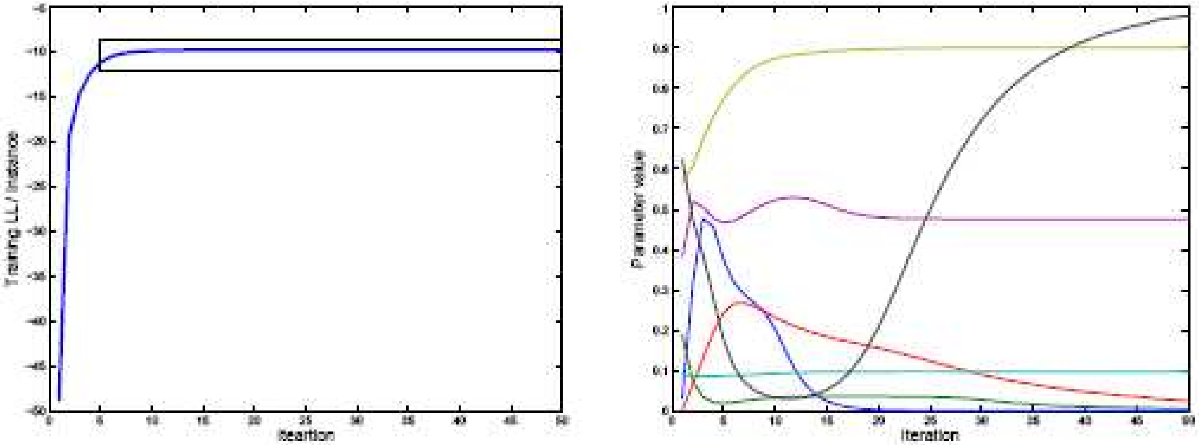
(a) (b)

Assessing convergence

- Can check whether parameters stop changing or LL stops changing. Can be quite different.
- Recall

$$\frac{\partial \ell(\boldsymbol{\theta} : \mathcal{D})}{\partial P(x \mid \boldsymbol{u})} = \frac{1}{P(x \mid \boldsymbol{u})} \sum_{m=1}^{M} P(x, \boldsymbol{u} \mid \boldsymbol{o}[m], \boldsymbol{\theta}).$$

- If p(x,u|o[m]) small, gradient is small, else O(M)
- Hence effects of param on LL can be small or large.

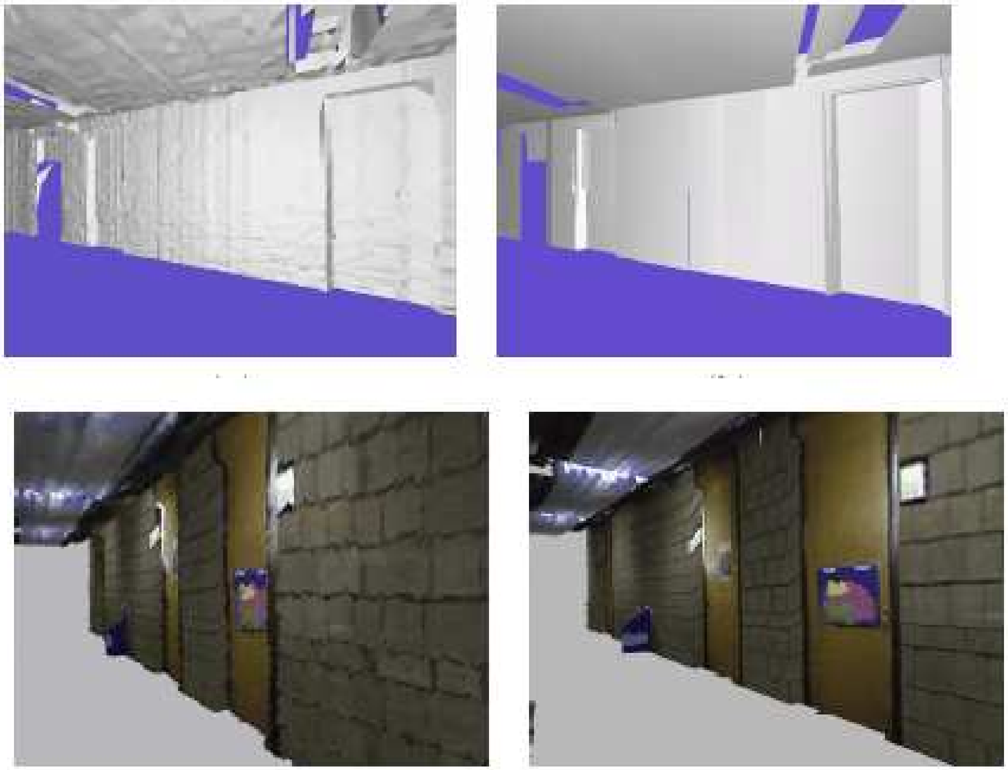


(a) (b)

Accelerating convergence

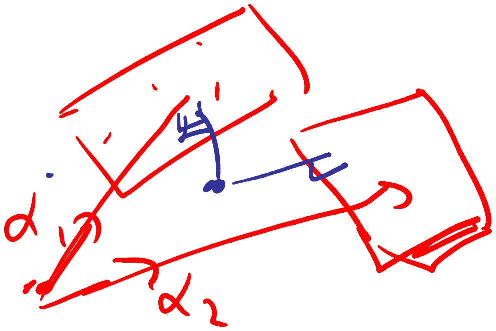
- Hard assignment EM (eg Kmeans). E step is searching over discrete assignments; this tends to converge faster (but to a worse solution).
- Hybrid EM/CG
- Over relaxation: step size > 1.
- Stochastic EM: since Q(H) = prod_m Q(hm|om), we can do inference on only a subset of the datacases (mini-batch) and then do an M step
- (Monte Carlo EM: sampling in the E step)

Example: fitting planes to 3d point clouds



$P(\mathbf{X}_m \mid C_m = k : \boldsymbol{\theta}_k)$ to be $\propto \mathcal{N}\left(d(x, p_k); 0, \sigma^2\right)$.

$$d(x, p_k) = |\alpha_k x - \beta_k|.$$



Variational EM

- Restrict Q distribution in E step to a tractable family, rather than p(H|D,theta)

$$\max_{\boldsymbol{\theta}} \max_{Q \in \mathcal{Q}} \Phi_{\mathcal{D}}[\boldsymbol{\theta}, Q]$$

- Eg do mean-field in the E step, then regular M step
- Maximizes a lower bound on the LL

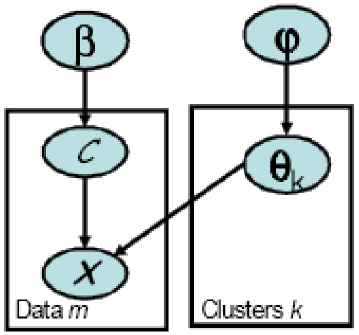
$$\ell(\boldsymbol{\theta} : \mathcal{D}) = \max_{Q} \Phi_{\mathcal{D}}[\boldsymbol{\theta}, Q] \geq \max_{Q \in \mathcal{Q}} \Phi_{\mathcal{D}}[\boldsymbol{\theta}, Q].$$

MCMC

- Can compute $p(\theta,H|D)$ using standard algorithms
- Parameter collapsed particles: sample θ, compute $p(H|D)$ analytically
- Data completion collapsed particles: sample H, compute $p(\theta|H,D)$ analytically

Marginalizing out H

- Bayesian Mixture model

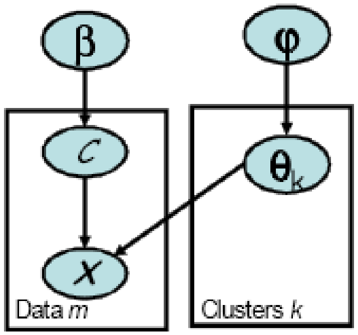


$$P(\mathcal{D} \mid \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, \boldsymbol{\lambda}) = \prod_{m=1}^{M} \left(\sum_{k=1}^{K} P(C[m] = c^k \mid \boldsymbol{\lambda}) P(\boldsymbol{x}[m] \mid C[m] = c^k, \boldsymbol{\theta}_k) \right)$$

$$P(\boldsymbol{\theta}_k \mid \boldsymbol{\lambda}, \boldsymbol{\theta}_{-k}, \mathcal{D}) \propto P(\mathcal{D} \mid \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, \boldsymbol{\lambda}).$$

Have to use MH

Marginalizing out \theta



Collapsed Gibbs sampling
Cf DP mixtures

$$P(\boldsymbol{\lambda} \mid \boldsymbol{c}) = Dirichlet(\alpha_0/K + |I_1(\boldsymbol{c})|, \ldots, \alpha_0/K + |I_K(\boldsymbol{c})|).$$

$$P(\boldsymbol{\theta}_k \mid \boldsymbol{c}, \mathcal{D}, \phi) = Q(\boldsymbol{\theta}_k \mid \mathcal{D}_{I_k(\boldsymbol{c})}, \phi) \propto P(\boldsymbol{\theta}_k \mid \phi) \prod_{m \in I_k(\boldsymbol{c})} P(\boldsymbol{x}[m] \mid \boldsymbol{\theta}_k),$$

$$P(C[m'] = k \mid \boldsymbol{c}_{-m'}, \mathcal{D}, \phi) \propto P(C[m'] = k \mid \boldsymbol{\lambda}, \boldsymbol{c}_{-m'}) P(\boldsymbol{x}[m'] \mid C[m'] = k, \boldsymbol{x}[I_k(\boldsymbol{c}_{-m'})], \phi).$$

$$P(C[m'] = k \mid \boldsymbol{c}_{-m'}, \mathcal{D}, \phi) \propto (|I_k(\boldsymbol{c}_{-m'})| + \alpha_0/K) Q(\boldsymbol{X} \mid \mathcal{D}_{I_k(\boldsymbol{c}_{-m'})}, \phi).$$

Variational Bayes

- Min KL(Q|P) where we assume

$$Q(\boldsymbol{\theta}, \mathcal{H}) = Q(\boldsymbol{\theta})Q(\mathcal{H}).$$

- Thm 19.3.6. If we have global param independence and Q(θ,H)=Q(θ) Q(H) then

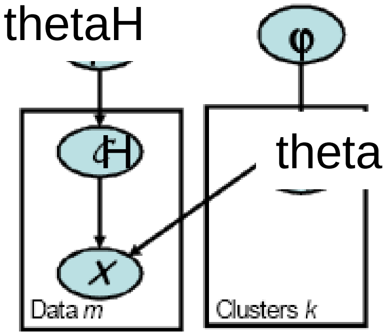
$$Q(\boldsymbol{\theta}, \mathcal{H}) = \prod_i Q(\boldsymbol{\theta}_{X_i|\boldsymbol{U}_i}) \prod_m Q(\boldsymbol{h}[m]).$$

- Hence we can optimize each Q(h[m]) separately – just like inference per case in the E step – and then optimize each Q(\theta_i) separately – just like optimizing each family in the M step
- E step: we do inference with expected params
- M step: we fit a distribution

VB for MixBernoulli

$$Q = Q(\boldsymbol{\theta}_H)\left[\prod_i Q(\boldsymbol{\theta}_{X_i|H})\right]\left[\prod_m Q(H[m])\right].$$

Beta Beta Bernoulli



thetaH

thetaXi

$$\begin{aligned}
Q(\boldsymbol{\theta}_H) &\propto \exp\left\{\ln P(\boldsymbol{\theta}_H) + \sum_m \boldsymbol{E}_{Q(H[m])}[\ln P(H[m] \mid \boldsymbol{\theta}_H)]\right\} \\
Q(\boldsymbol{\theta}_{X_i|H}) &\propto \exp\left\{\ln P(\boldsymbol{\theta}_{X_i|H}) + \sum_m \boldsymbol{E}_{Q(H[m])}\left[\ln P(x_i[m] \mid H[m], \boldsymbol{\theta}_{X_i|H})\right]\right\} \\
Q(H[m]) &\propto \exp\Big\{\boldsymbol{E}_{Q(\boldsymbol{\theta}_H)}[\ln P(H[m] \mid \boldsymbol{\theta}_H)] \\
&\qquad + \sum_i \boldsymbol{E}_{Q(\boldsymbol{\theta}_{X_i|H})}[\ln P(x_i[m] \mid H[m], \boldsymbol{\theta}_{X_i|H})]\Big\}.
\end{aligned}$$

$$\ln P(\boldsymbol{\theta}_H = \langle \theta_{h^0}, \theta_{h^1}\rangle) = \ln c + (\alpha_{h^0} - 1)\ln\theta_{h^0} + (\alpha_{h^1} - 1)\ln\theta_{h^1}.$$

$$\boldsymbol{E}_{Q(H[m])}[\ln P(H[m] \mid \boldsymbol{\theta}_H = \langle\theta_{h^0}, \theta_{h^1}\rangle)] = Q(H[m] = h^0)\ln\theta_{h^0} + Q(H[m] = h^1)\ln\theta_{h^1}.$$

$$\begin{aligned}
Q(\boldsymbol{\theta}_H = \langle\theta_{h^0}, \theta_{h^1}\rangle) &\propto \exp\left\{\left(\alpha_{h^0} + \sum_m Q(H[m] = h^0) - 1\right)\ln\theta_{h^0} + \right. \\
&\qquad \left.\left(\alpha_{h^1} + \sum_m Q(H[m] = h^1) - 1\right)\ln\theta_{h^1}\right\} \\
&= \theta_{h^0}^{\alpha_{h^0} + \sum_m Q(H[m]=h^0) - 1}\,\theta_{h^1}^{\alpha_{h^1} + \sum_m Q(H[m]=h^1) - 1}.
\end{aligned}$$

$$\begin{aligned}
\alpha'_{h^0} &= \alpha_{h^0} + \sum_m Q(H[m] = h^0) \\
\alpha'_{h^1} &= \alpha_{h^1} + \sum_m Q(H[m] = h^1).
\end{aligned}$$

VB update for H[m]

Regular E step

$$P(H[m] \mid x_1[m], \ldots x_n[m]) \propto P(H[m] \mid \boldsymbol{\theta}_H) \prod_i P(x_i[m] \mid H[m], \boldsymbol{\theta}_{X_i|H}).$$

VB version

$$\boldsymbol{E}_{Q(\boldsymbol{\theta}_{X_i|H})}[\ln P(x_i \mid H[m], \boldsymbol{\theta}_{X_i|H})] = \int_0^1 Q(\theta_{x_i|H[m]}) \ln \theta_{x_i|H[m]} d\theta_{x_i|H[m]}.$$

$$\boldsymbol{E}_{Q(\boldsymbol{\theta}_{X_i|H})}[\ln P(x_i \mid H[m], \boldsymbol{\theta}_{X_i|H})] = \varphi(\alpha'_{x_i|h}) - \varphi(\sum_{x'_i} \alpha'_{x'_i|h})$$

where α' *are the hyperparameters of the posterior approximation in* $Q(\boldsymbol{\theta}_{X_i|H})$ *and* $\varphi(z) = (\ln \Gamma(z))' = \frac{\Gamma'(z)}{\Gamma(z)}$ *is the* digamma *function, which is equal to* $\ln(z)$ *plus a polynomial function of* $\frac{1}{z}$. *And so, for* $z \gg 1$, $\varphi(z) \approx \ln(z)$. *Using this approximation, we see that*

$$\boldsymbol{E}_{Q(\boldsymbol{\theta}_{X_i|H})}[\ln P(x_i \mid H[m], \boldsymbol{\theta}_{X_i|H})] \approx \ln \frac{\alpha'_{x_i|h}}{\sum_{x'_i} \alpha'_{x'_i|h}},$$

Variational methods

- From Lecture 10:
- Minimize

$$\begin{aligned} D(Q||P) &= \ln Z - F(\tilde{P}, Q) \\ F(\tilde{P}, Q) &\overset{\text{def}}{=} H_Q(x) + \sum_i E_{C_i \sim Q} \ln \psi_i(C_i) \end{aligned}$$

- This always increases the lower bound and will always converge

Mean field approximation

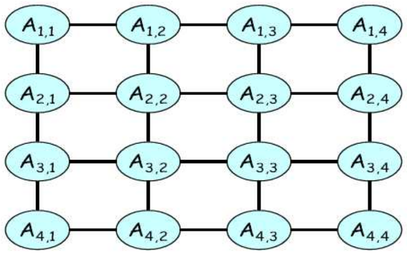
- Let us assume the approximate posterior is fully factorized

$$Q(x) = \prod_i Q_i(x_i)$$

- Then the objective (negative free energy) is

$$\begin{aligned} F(\tilde{P}, Q) &\stackrel{\text{def}}{=} H_Q(x) + \sum_c E_{X_c \sim Q} \ln \phi_c(X_c) \\ &= \sum_i H(Q_i) + \sum_c \sum_{x_c} (\prod_{i \in c} Q_i(x_{c,i})) \ln \phi_c(x_c) \end{aligned}$$

- Eg 4x4 grid $O(n_e K^2)$ for energy, $O(n_e K)$ for H



$$\begin{aligned} F[\tilde{P}_\Phi, Q] \;&=\; E_{\{A_{1,1},A_{2,1}\}\sim Q}[\ln \phi(A_{1,1}, A_{2,1})] + E_{\{A_{2,1},A_{3,1}\}\sim Q}[\ln \phi(A_{2,1}, A_{3,1})] + E_{\{A_{3,1},A_{4,1}\}\sim Q} \\ &\quad \ldots \\ &\quad E_Q[\ln \phi(A_{1,1}, A_{1,2})] + E_Q[\ln \phi(A_{1,2}, A_{1,3})] + E_Q[\ln \phi(A_{1,3}, A_{1,4})] + \\ &\quad \ldots \\ &\quad H_Q(A_{1,1}) + H_Q(A_{1,2}) + H_Q(A_{1,3}) + H_Q(A_{1,4}) + \\ &\quad \ldots \\ &\quad H_Q(A_{4,1}) + H_Q(A_{4,2}) + H_Q(A_{4,3}) + H_Q(A_{4,4}) \end{aligned}$$

Convexity

- Objective is concave in each arg (entropy is concave in each Q_i, expected energy is linear in Q_i)

$$F(\tilde{P}, Q) \quad = \quad \sum_i H(Q_i) + \sum_c \sum_{x_c} (\prod_{i \in c} Q_i(x_{c,i})) \ln \phi_c(x_c)$$

- The set of completely factorized distributions is not convex

$$Q^3(x) = \lambda \prod_i Q^1(x_i) + (1-\lambda) \prod_i Q_i^2(x_i)$$

Not factorized

- Hence we are optimizing the objective over a non-convex space, and will be subject to local maxima
- Let us derive equations that characterize the fixed points. These could correspond to saddle points or local minima, but such points are unstable and unlikely to be the result of our iterative update scheme.

Notation

- Define

$$
\begin{aligned}
\langle f(x_h) \rangle &\stackrel{\text{def}}{=} \sum_{x_h} \left[\prod_{i \in h} Q_i(x_i) \right] f(x_h) \\
\langle f(x_h) \rangle_{j,k} &\stackrel{\text{def}}{=} \sum_{x_h \backslash x_j} \left[\prod_{i \in h, i \neq j} Q_i(x_i) \right] f(x_h | x_j = k) \\
\langle f(x_h) \rangle &= \sum_k Q_j(x_j = k) \langle f(x_h) \rangle_{j,k} \\
\ln p(x_v) &\geq \sum_c \langle \ln \phi_c(x_c) \rangle + \sum_i H(Q_i) \\
&= \sum_k Q_j(k) \sum_c \langle \ln \phi_c(x_c) \rangle_{j,k} + H(Q_j) + \sum_{i \neq j} H(Q_i)
\end{aligned}
$$

We mostly follow Tommi Jaakkola's notation rather than Daphne Koller's

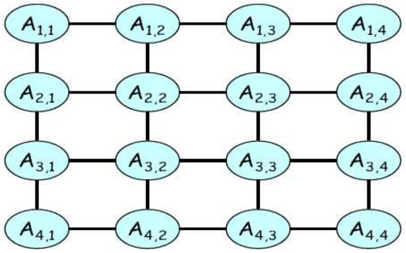
Mean field equations

$$\begin{aligned}
\ln p(x_v) &\geq \sum_k Q_j(k) \sum_c \langle \ln \phi_c(x_c) \rangle_{j,k} + H(Q_j) + \sum_{i \neq j} H(Q_i) \\
&\overset{\text{def}}{=} L(Q_j) \\
S_{j,k} &\overset{\text{def}}{=} \sum_{c:j \in c} \langle \ln \phi_c(x_c) \rangle_{j,k} \\
L(Q_j) &= \sum_k Q_j(k)(S_{j,k} - \ln Q_j(k)) + C \\
L(Q_j, \lambda) &\overset{\text{def}}{=} L(Q_j) + \lambda(\sum_{k'} Q_j(k') - 1) \\
\frac{\partial}{\partial Q_j(k)} L(Q_j, \lambda) &= S_{j,k} - \ln Q_j(k) - 1 + \lambda = 0 \\
Q_j(k) &= \exp(S_{j,k}) \exp(\lambda - 1) \\
&= \frac{1}{Z_j} \exp(\sum_c \langle \ln \phi_c(x_c) \rangle_{j,k})
\end{aligned}$$

Example: grid

$$Q(x_i) = \frac{1}{Z_i} \exp \left\{ \sum_{\phi : X_i \in Scope[\phi]} E_{(\boldsymbol{U}_\phi - \{X_i\}) \sim Q}[\ln \phi(\boldsymbol{U}_\phi, x_i)] \right\}$$

$$Q(a_{i,j}) = \frac{1}{Z_{i,j}} \exp \left\{ \begin{array}{l} \sum_{a_{i-1,j}} Q(a_{i-1,j}) \ln(\phi(a_{i-1,j}, a_{i,j})) + \\ \sum_{a_{i,j-1}} Q(a_{i,j-1}) \ln(\phi(a_{i,j-1}, a_{i,j})) + \\ \sum_{a_{i+1,j}} Q(a_{i+1,j}) \ln(\phi(a_{i,j}, a_{i+1,j})) + \\ \sum_{a_{i,j+1}} Q(a_{i,j+1}) \ln(\phi(a_{i,j}, a_{i,j+1})) \end{array} \right\}.$$



EM

- Suppose we want to find a MAP estimate

$$\max_{\theta} \log p(\theta) + \sum_{n} \log p(x_n|\theta)$$

- If we have latent variables Z we can use EM
- E step: compute expected complete data log joint

$$f(\theta, \theta_{old}) \quad = \quad \log p(\theta) + \sum_{n=1}^{N} \sum_{z} p(z|x_n, \theta_{old}) \log p(z, x_n|\theta)$$

- M step: set

$$\theta_{new} = \arg\max f(\theta, \theta_{old})$$

Variational EM

- Consider the negative free energy

$$F(x, Q, \theta) = \sum_z Q(z) \log p(x, z|\theta) + H(Q)$$

- Earlier we showed this is a lower bound on the log-likelihood

$$\begin{aligned} F(x, Q, \theta) &= \ln Z(x, \theta) - D(Q||p(z|x, \theta)) \\ \log p(x|\theta) &= \ln Z = \max_Q F(x, Q, \theta) = F(x, Q^*, \theta) \geq F(x, Q, \theta) \end{aligned}$$

- Where the bound is tight if $Q^*(z) = p(z|x, \theta)$
- E step: find Q_n(z) that maximize

$$F(x_n, Q_n, \theta_{old})$$

- M step: find \theta that maximize

$$\log p(\theta) + \sum_n F(x_n, Q_n, \theta)$$

Variational EM

- An exact E step is equivalent to setting

$$Q_n(z) = p(z|x_n, \theta_{old})$$

- The corresponding M step maximizes

$$\begin{aligned}\sum_n F(x_n, Q_n, \theta) &= \sum_n [\sum_z p(z|x_n, \theta_{old}) \log p(z, x_n|\theta)] + H(Q_n) \\ &= f(\theta, \theta_{old}) + \sum_n H(Q_n)\end{aligned}$$

- Since $H(Q_n)$ is independent of θ, this reduces to the standard EM algorithm.
- Generalized EM merely increases (not maximizes) θ in the M step.
- Similarly we can simply improve Q_n in the E step

Neal and Hinton, “A new view of the EM algorithm”, 1998

Variational Bayes

- We can replace the point estimate of θ with a distribution and try to minimize

$$D(Q(z_{1:N},\theta|x_{1:N})||p(z_{1:N},\theta|x_{1:N}))$$

- The distinction between E and M vanishes: we are just doing sequential updates of Q(Z_n) and Q(θ)
- This gives us the benefits of being Bayesian for the same computational speed as EM

VB for univariate Gaussian

$$q_j^\star(\mathbf{Z}_j) = \frac{\exp\left(\mathbb{E}_{i\neq j}[\ln p(\mathbf{X},\mathbf{Z})]\right)}{\displaystyle\int \exp\left(\mathbb{E}_{i\neq j}[\ln p(\mathbf{X},\mathbf{Z})]\right)\,\mathrm{d}\mathbf{Z}_j}.$$

$$\ln q_j^\star(\mathbf{Z}_j) = \mathbb{E}_{i\neq j}[\ln p(\mathbf{X},\mathbf{Z})] + \text{const.}$$

$$p(\mathcal{D}|\mu,\tau) = \left(\frac{\tau}{2\pi}\right)^{N/2} \exp\left\{-\frac{\tau}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}.$$

$$\begin{aligned} p(\mu|\tau) &= \mathcal{N}\left(\mu|\mu_0,(\lambda_0\tau)^{-1}\right) \\ p(\tau) &= \mathrm{Gam}(\tau|a_0,b_0) \end{aligned}$$

$$q(\mu,\tau) = q_\mu(\mu)q_\tau(\tau).$$

$$\begin{aligned} \ln q_\mu^\star(\mu) &= \mathbb{E}_\tau\left[\ln p(\mathcal{D}|\mu,\tau) + \ln p(\mu|\tau)\right] + \text{const.} \\ &= -\frac{\mathbb{E}[\tau]}{2}\left\{\lambda_0(\mu-\mu_0)^2 + \sum_{n=1}^{N}(x_n-\mu)^2\right\} + \text{const.} \end{aligned}$$

Gaussian

$$\begin{aligned} \mu_N &= \frac{\lambda_0\mu_0 + N\overline{x}}{\lambda_0 + N} \\ \lambda_N &= (\lambda_0 + N)\mathbb{E}[\tau]. \end{aligned}$$

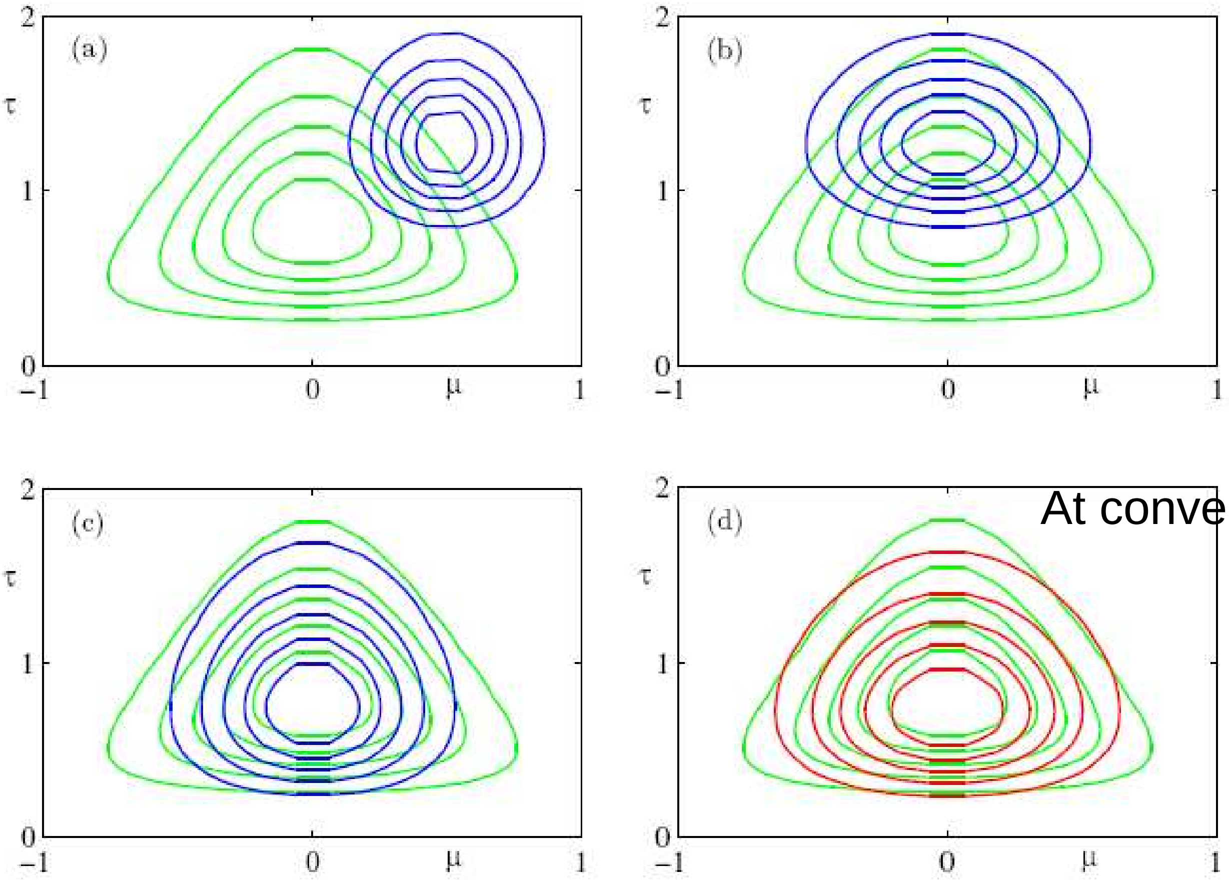
$$\begin{aligned} \ln q_\tau^\star(\tau) &= \mathbb{E}_\mu\left[\ln p(\mathcal{D}|\mu,\tau) + \ln p(\mu|\tau)\right] + \ln p(\tau) + \text{const.} \\ &= (a_0-1)\ln\tau - b_0\tau + \frac{N+1}{2}\ln\tau \\ &\quad -\frac{\tau}{2}\mathbb{E}_\mu\left[\sum_{n=1}^{N}(x_n-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right] + \text{const.} \end{aligned}$$

Gamma

$$\begin{aligned} a_N &= a_0 + \frac{N+1}{2} \\ b_N &= b_0 + \frac{1}{2}\mathbb{E}_\mu\left[\sum_n (x_n-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right]. \end{aligned}$$

Bishop p471

VB for univariate Gaussian



At convergence

Green = exact posterior (NormalGamma), blue = factorized approximation

VB for mixtures of Gaussians

Inference

$$q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\mathbf{Z})q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}).$$

$$\ln q^\star(\mathbf{Z}) = \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Lambda}}[\ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] + \text{const.}$$

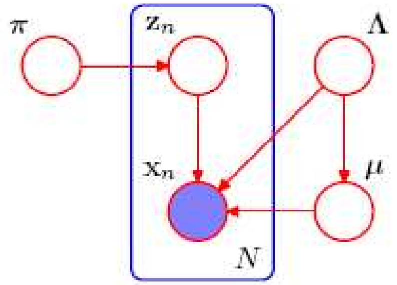
$$\ln q^\star(\mathbf{Z}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \ln \rho_{nk} + \text{const.}$$

$$\begin{aligned}\ln \rho_{nk} \;=\; & \mathbb{E}[\ln \pi_k] + \frac{1}{2}\mathbb{E}\left[\ln |\boldsymbol{\Lambda}_k|\right] - \frac{D}{2}\ln(2\pi) \\ & -\frac{1}{2}\mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k}\left[(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Lambda}_k(\mathbf{x}_n - \boldsymbol{\mu}_k)\right]\end{aligned}$$

$$q^\star(\mathbf{Z}) \propto \prod_{n=1}^{N}\prod_{k=1}^{K} \rho_{nk}^{z_{nk}}.$$

Multinomial (soft responsibilities), as in EM,
except we used expected parameters rather than plug-in

Model



$$p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})p(\mathbf{Z}|\boldsymbol{\pi})p(\boldsymbol{\pi})p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$$

$$p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right)^{z_{nk}}$$

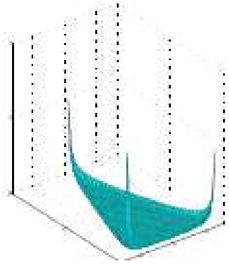
$$p(\mathbf{Z}|\boldsymbol{\pi}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k^{z_{nk}}.$$

$$p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}_0) = C(\boldsymbol{\alpha}_0)\prod_{k=1}^{K}\pi_k^{\alpha_0 - 1}$$

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{k=1}^{K}\mathcal{N}\left(\boldsymbol{\mu}_k|\mathbf{m}_0, (\beta_0\boldsymbol{\Lambda}_k)^{-1}\right)\mathcal{W}(\boldsymbol{\Lambda}_k|\mathbf{W}_0, \nu_0)$$

Automatic model selection

- Recall π ~ Dir(α). If α << 1, we prefers skewed π and hence sparse z.



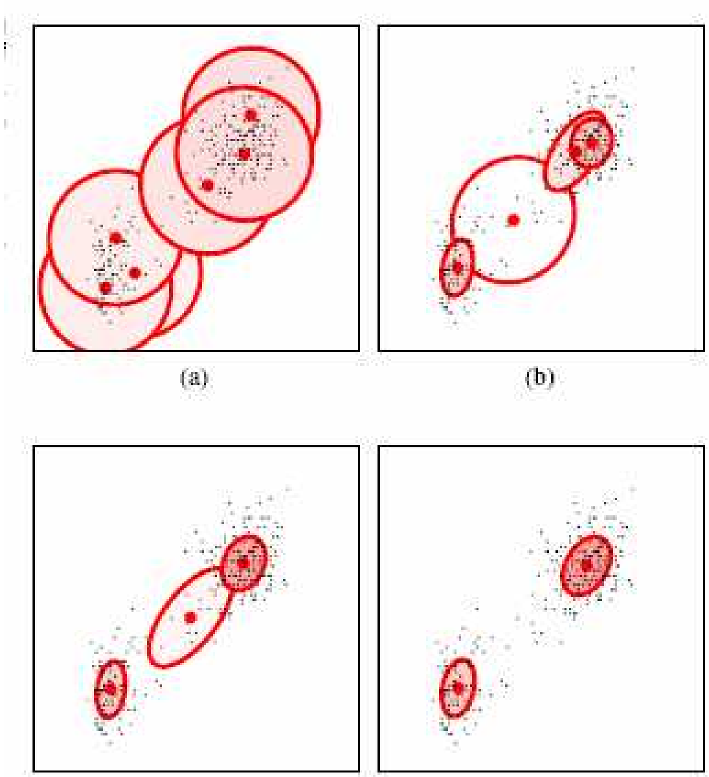
- MAP estimate from regular EM is

$$\hat{\pi}_k \quad = \quad \frac{\sum_n r_{nk} + \alpha_k - 1}{\sum_k (r_{nk} + \alpha_k - 1)} = \frac{N_k + \alpha - 1}{N + K\alpha - K}$$

- Posterior mean estimate from VB is

$$\hat{\pi}_k \quad = \quad \frac{\sum_n r_{nk} + \alpha_k}{\sum_k (r_{nk} + \alpha_k)} = \frac{N_k + \alpha}{N + K\alpha} \rightarrow \frac{\alpha}{N + K\alpha} \rightarrow 0$$

Selecting K with one run of VB



Variational message passing

- Consider a DAG model

$$p(\mathbf{x}) = \prod_i p(\mathbf{x}_i|\mathrm{pa}_i)$$

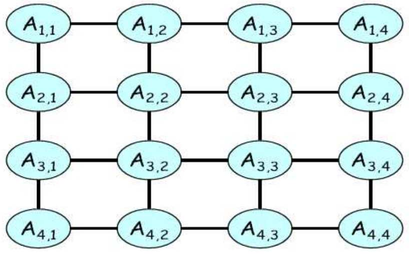
- The mean field equations are

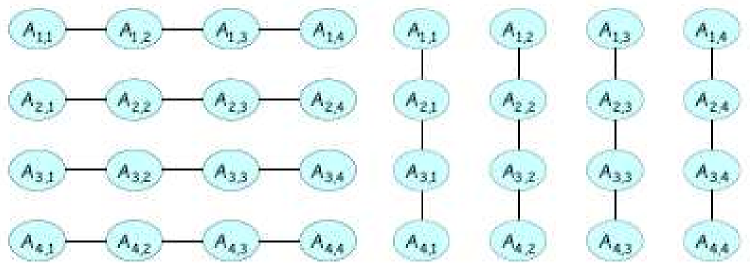
$$\ln q_j^\star(\mathbf{x}_j) = \mathbb{E}_{i\neq j}\left[\sum_i \ln p(\mathbf{x}_i|\mathrm{pa}_i)\right] + \mathrm{const.}$$

- The only terms that depend on x_j are in x_j's Markov blanket
- If all CPDs have conjugate-exponential form, the VB updates can be converted into a msg passing algorithm
- VIBES software (John Winn)

Structured variational approx

- Rather than assuming Q is fully factorized, we can use any structure for which computing the expectations of ln ϕ_c and the entropy is tractable


$A_{1,1}$ $A_{1,2}$ $A_{1,3}$ $A_{1,4}$
$A_{2,1}$ $A_{2,2}$ $A_{2,3}$ $A_{2,4}$
$A_{3,1}$ $A_{3,2}$ $A_{3,3}$ $A_{3,4}$
$A_{4,1}$ $A_{4,2}$ $A_{4,3}$ $A_{4,4}$


$A_{1,1}$ $A_{1,2}$ $A_{1,3}$ $A_{1,4}$ $A_{1,1}$ $A_{1,2}$ $A_{1,3}$ $A_{1,4}$
$A_{2,1}$ $A_{2,2}$ $A_{2,3}$ $A_{2,4}$ $A_{2,1}$ $A_{2,2}$ $A_{2,3}$ $A_{2,4}$
$A_{3,1}$ $A_{3,2}$ $A_{3,3}$ $A_{3,4}$ $A_{3,1}$ $A_{3,2}$ $A_{3,3}$ $A_{3,4}$
$A_{4,1}$ $A_{4,2}$ $A_{4,3}$ $A_{4,4}$ $A_{4,1}$ $A_{4,2}$ $A_{4,3}$ $A_{4,4}$

$$Q(\mathcal{X}) = \frac{1}{Z_Q} \prod_{j=1}^{J} \psi_j$$

ϕ = model, ψ = approx

Corollary 11.5.13: *If* $Q(\mathcal{X}) = \frac{1}{Z_Q} \prod_j \psi_j$*, then the potential* ψ_j *is a stationary point of the energy functional if and only if:*

$$\psi_j(\boldsymbol{c}_j) \propto \exp\left\{ \boldsymbol{E}_Q\left[\ln \tilde{P}_\Phi \mid \boldsymbol{c}_j\right] - \sum_{k \neq j} \boldsymbol{E}_Q[\ln \psi_k \mid \boldsymbol{c}_j] \right\}. \qquad (11.59)$$