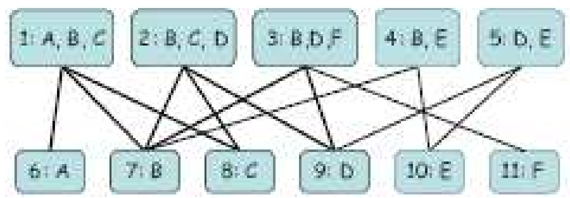
Stat 521A
Lecture 10

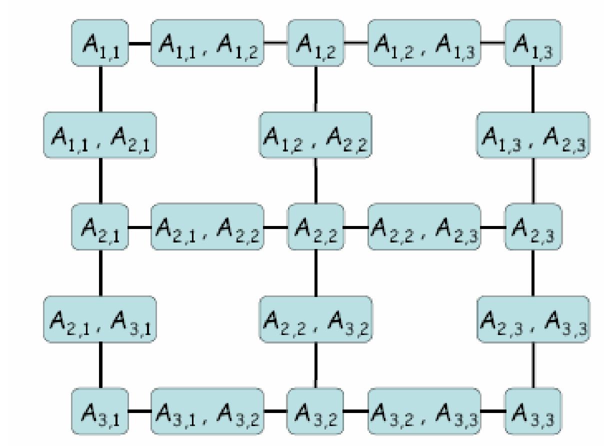
Outline

- Belief propagation: entropy approximations (11.3.7)
- Expectation propagation (11.4)
- Mean field (11.5.1)
- Variational EM/ Bayes (Bishop 10.1-10.2)
- Structured variational (11.5.2)

Bethe cluster graphs

- Suppose we create one cluster for each original factor, and one cluster for each node.

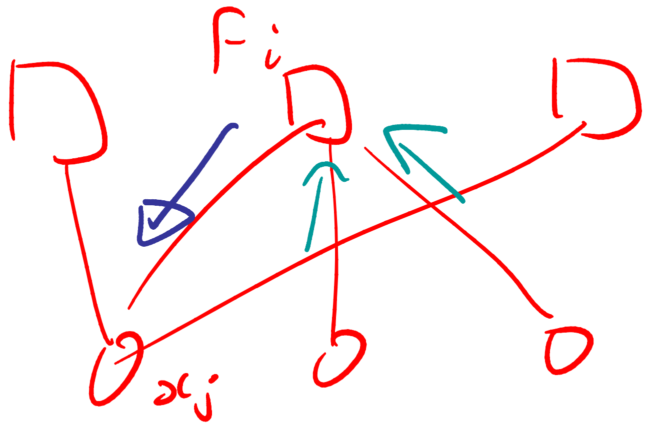




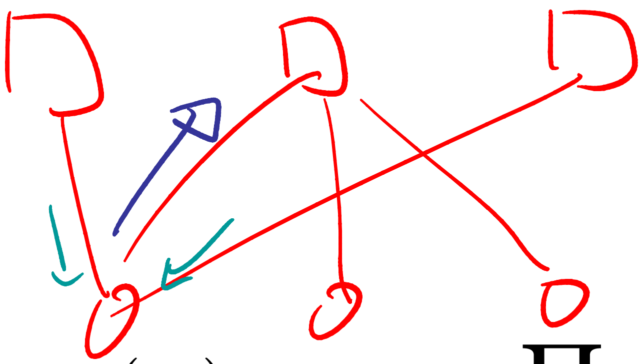
- Then for a pairwise MRF, propagating $C_i - C_{ij} - C_j$ is equivalent to sending msgs from node i to node j via edge ij.
- In general, BP on the Bethe CG = BP on the factor graph.

BP on factor graphs

Bishop p406



$$\mu_{f_i \to x_j}(x_j) \quad = \quad \sum_{c_i \setminus x_j} f(c_i) \prod_{k \in nb(f_i) \setminus x_j} \mu_{x_k \to f_i}(x_k)$$



$$\mu_{x_i \to f_j}(x_i) \quad = \quad \prod_{k \in nb(x_i) \setminus f_j} \mu_{f_k \to x_i}(x_i)$$

Bethe approximation to entropy

- Thm 11.3.10. If Q is a calibrated set of beliefs for a Bethe approximation CG then the factored energy is given by

$$
\begin{aligned}
\tilde{F}(\tilde{P}, Q) &\stackrel{\text{def}}{=} \sum_{\phi} E_{\beta_\phi} \ln \phi + \sum_{\phi} H_{\beta_\phi}(C_\phi) - \sum_{s} H_{\mu_s}(S_s) \\
&= \sum_{\phi} E_{\beta_\phi} \ln \phi + \sum_{\phi} H_{\beta_\phi}(C_\phi) - \sum_{i} (d_i - 1) H_{\beta_i}(X_i)
\end{aligned}
$$

  where d_i = #factors that contain Xi.

- If Xi appears in di factors, by RIP, it appears in (di-1) sepsets. Hence we count the entropy of each Xi once in total.

Weighted approximation to entropy

- Consider a cluster graph, each of whose clusters (regions) has a counting number μ_r. Define the weighted approximate entropy as

$$H_Q^{\mu}(X) = \sum_r \mu_r H_{\beta_r}(C_r)$$

- For a Bethe-structured CG, we set

$$\mu_i = 1 - \sum_{r \in nb_i} \mu_r$$

- If we set μ_r=1, we recover the Bethe approximation.
- Let us consider more general weightings.

Convex approximation to entropy

- Def 11.3.13. We say that μ_r are convex counting numbers if there exist non-negative numbers ν_r, ν_i, $\nu_{r,i}$ st

$$
\begin{aligned}
\mu_r &= \nu_r + \textstyle\sum_{i\,:\,X_i \in \boldsymbol{C}_r} \nu_{r,i} \quad \textit{for all } r \\
\mu_i &= \nu_i - \textstyle\sum_{r\,:\,X_i \in \boldsymbol{C}_r} \nu_{r,i} \quad \textit{for all } i
\end{aligned}
$$

- Then

$$
\sum_r \mu_r \mathbb{H}_{\beta_r}(C_r) + \sum_i \mu_i \mathbb{H}_{\beta_i}(X_i) = \sum_r \nu_r \mathbb{H}_{\beta_r}(C_r) + \sum_{r, X_i \in \boldsymbol{C}_r} \nu_{r,i}(\mathbb{H}_{\beta_r}(C_r) - \mathbb{H}_{\beta_i}(X_i)) \sum_i \nu_i \mathbb{H}_{\beta_i}(X_i)
$$

- Thm 11.3.14. The above eqn is concave for any set of beliefs Q which satisfy marginal consistency constraints.

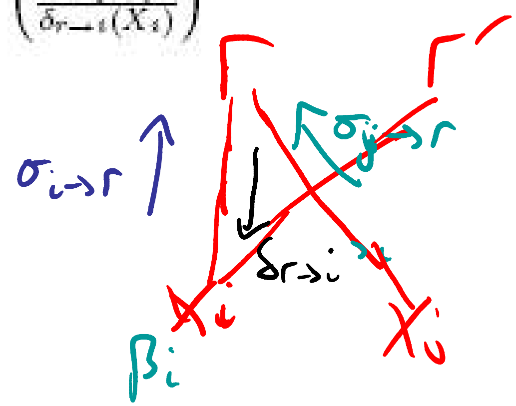
Convex BP

Algorithm 11.2 Convergent message passing algorithm for Bethe-structured region graphs with convex counting numbers

Procedure Convex-BP-Msg (
  $\psi_r[\boldsymbol{C}_r]$ // set of initial potentials
  $\sigma_{i\to r}(\boldsymbol{C}_r)$ // Current node-to-region messages
)

for $i = 1, \ldots, n$
  // Compute incoming messages from neighboring regions to X_i
  for $r \in \mathrm{Nb}_i$
    $\delta_{r\to i}(X_i) \leftarrow \sum_{\boldsymbol{C}_r - X_i} \left(\psi_r[\boldsymbol{C}_r] \prod_{j \in \mathrm{Nb}_r - \{i\}} \sigma_{j\to r}(\boldsymbol{C}_r) \right)^{\frac{1}{\hat{\nu}_{i,r}}}$
  // Compute beliefs for X_i, renormalizing to avoid numerical underflows
  $\beta_i[X_i] \leftarrow \propto \prod_{r \in \mathrm{Nb}_i} (\delta_{r\to i}(X_i))^{\hat{\nu}_{i,r}/\hat{\nu}_i}$
  // Compute outgoing messages from X_i to neighboring regions
  for $r \in \mathrm{Nb}_i$
    $\sigma_{i\to r}(\boldsymbol{C}_r) \leftarrow \left(\psi_r[\boldsymbol{C}_r] \prod_{j \in \mathrm{Nb}_r - \{i\}} \sigma_{j\to r}(\boldsymbol{C}_r) \right)^{-\frac{\nu_{i,r}}{\hat{\nu}_{i,r}}} \left(\frac{\beta_i[X_i]}{\delta_{r\to i}(X_i)} \right)^{\nu_r}$
return $\{\sigma_{i\to r}(\boldsymbol{C}_r)\}_{i, r \in \mathrm{Nb}_i}$

$$\hat{\nu}_i = \nu_i + \sum_{r \in \mathrm{Nb}_i} \nu_r; \qquad \hat{\nu}_{i,r} = \nu_r + \nu_{i,r}.$$



TRW

- Tree reweighting algorithm (TRW) uses the following convex counting numbers, given a distribution over trees T st each edge in the pairwise network is present in at least 1 tree

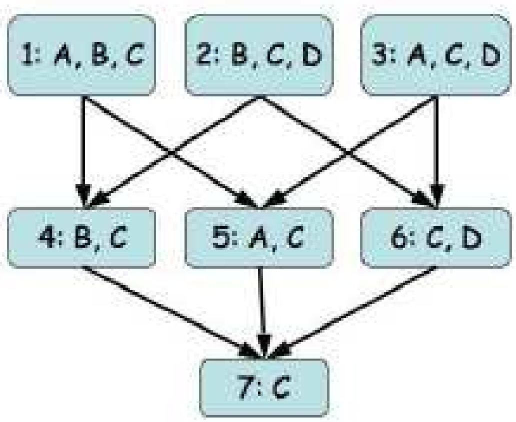
$$\begin{aligned} \mu_i &= -\sum_{T \ni X_i} \rho(T) \\ \mu_{i,j} &= \sum_{T \ni (X_i, X_j)} \rho(T) \end{aligned}$$

Convex or not?

- When standard BP converges, the Bethe approximation to the entropy is often more accurate than the convex approximation.
- However, it is desirable to have a convex inference engine in the inner loop of learning.
- If you train with a convex approximation, there are some arguments you should use the same convex approx at test time for decoding.

Region graphs (11.3.7.3)

- One can use more general CGs than the Bethe construction, which lets you model higher order interactions which are intermediate between the original factors and singletons.
- Resulting algorithm is complex.



Approximate messages

- Suppose we use a cluster tree (or graph), but approximate the messages eg. to prevent them becoming “too fat”
- If the clusters have internal structure, we can efficiently combine factored incoming messages with factored clusters to get factored outgoing messages
- We can also use this to combine non conjugate distributions: eg we approximate a non-conjugate likelihood by a simple form (eg MVN) and combine with a simple cluster potential (eg MVN) to get a simple posterior for the next step

Assumed density filtering (ADF)

- Consider sequential Bayesian updating in which we assume the prior $p(\theta_{t-1}|y_{1:t-1})$ lives in some tractable family Q (eg MVN).
- At each step, we do 1 step of Bayesian updating to get the posterior $p(\theta_t|y_{1:t})$ and then do an M-projection to get the best approximation within Q (=moment matching for linear exp fam)

$$\begin{array}{ccc} \theta_{t-1} & \to & \theta_t \\ \downarrow & & \downarrow \\ y_{t-1} & & y_t \end{array}$$

$$\tilde{s}_{t-1} \in Q \xrightarrow{\text{update}(y_t)} s_t \xrightarrow{\text{project}} \tilde{s}_t$$

Eg y_t|\theta$_t$ = logistic(y_t x_t^T θ_t), $\theta_t=\theta_{t-1}$

Eg. y_t|theta$_t$ = Gauss, theta$_t$ | theta$_{t-1}$ = mix Gauss

ADF cont'd

- We combine msg from past (prior) with local evidence (likelihood), project, then compute new msg

$$
\begin{aligned}
b_{t-1,t} &\propto \phi_{t-1,t}\mu_{t-1} \\
\tilde{b}_{t-1,t} &= \text{proj}(b_{t-1,t}, Q) \\
\mu_t &= \sum_t \tilde{b}_{t-1,t}
\end{aligned}
$$

Expectation propagation

- For batch problems, ADF is suboptimal, and depends on order of data.
- EP idea: add backwards pass

$$
\begin{aligned}
b^*_{t,t+1} &= \frac{\tilde{b}_{t,t+1}}{\mu_{t+1}} \mu^*_{t+1} \\
\tilde{b}^*_{t,t+1} &= \text{proj}(b^*_{t,t+1}, Q) \\
\mu^*_t &= \sum_{t+1} \tilde{b}^*_{t,t+1}
\end{aligned}
$$

- Since msgs no longer exact, need to iterate

Division = subtraction of natural params

- Assume all beliefs and msgs are linear exponential families. Then

$$\tilde{\delta}_{i\to j} = \frac{\tilde{\sigma}_{i\to j}}{\tilde{\delta}_{j\to i}} \propto \exp\left\{ \langle (\theta_{\tilde{\sigma}_{i\to j}} - \theta_{\tilde{\delta}_{j\to i}}), \tau_{i,j}(s_{i,j}) \rangle \right\}$$

- This can result in negative values for the natural params (eg Gaussians with –ve variance).
- But undirected GMs with tabular potentials are in the linear exp family and can always be used to represent valid beliefs/msgs

Projection

- How compute natural parameters of a msg?
- Compute the expected statistics of the separator, according to the current approximate beliefs

$$\theta_{\tilde{\delta}_{i \to j}} \leftarrow \text{M-project}_{i,j}(E_{\boldsymbol{S}_{i,j} \sim \tilde{\beta}_i}[\tau_{i,j}(\boldsymbol{S}_{i,j})]) - \theta_{\tilde{\delta}_{j \to i}}.$$

- Computing the expectation can be made tractable if β_i has factored structure.
- In general, the M projection can be hard.
- But if we have discrete variables, and Q is fully factorized, it amounts to computing a product of marginals.

Variational analysis

- We optimize the same (approximate) objective as before (factored free energy), but relax the local consistency conditions so we only match statistics (eg marginals) instead of full distributions

EP-Optimize

Find Q

that maximize $\tilde{F}[\tilde{P}_\Phi, Q]$

subject to

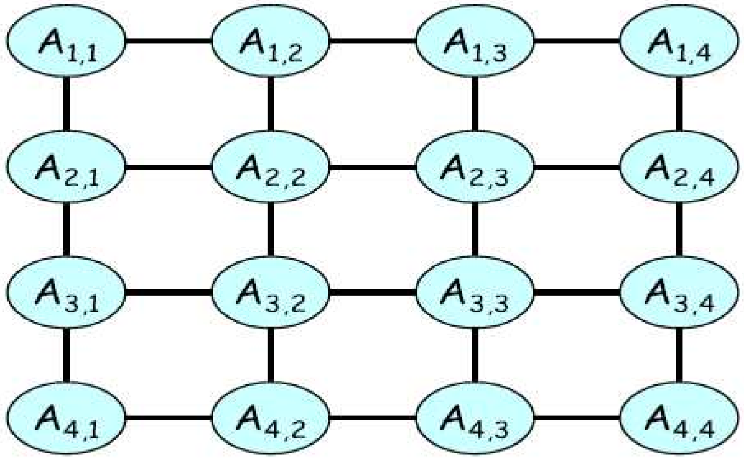
$$
\begin{array}{rcll}
E_{\boldsymbol{s}_{i,j} \sim \mu_{i,j}}[\tau_{i,j}] & = & E_{\boldsymbol{s}_{i,j} \sim \beta_j}[\tau_{i,j}] & \forall (i\!-\!j) \in \mathcal{E}_\mathcal{T} \\
\sum_{\boldsymbol{c}_i} \beta_i[\boldsymbol{c}_i] & = & 1 & \forall i \in \mathcal{V}_\mathcal{T} \\
\sum_{\boldsymbol{s}_{i,j}} \mu_{i,j}[\boldsymbol{s}_{i,j}] & = & 1 & \forall (i\!-\!j) \in \mathcal{E}_\mathcal{T} \\
\beta_i[\boldsymbol{c}_i] & \geq & 0 & \forall i \in \mathcal{V}_\mathcal{T}, \boldsymbol{c}_i \in Val(\boldsymbol{C}_i)
\end{array}
$$

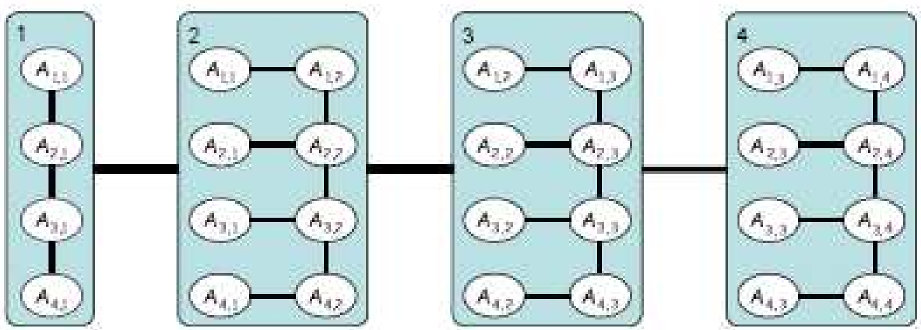
EP msg passing

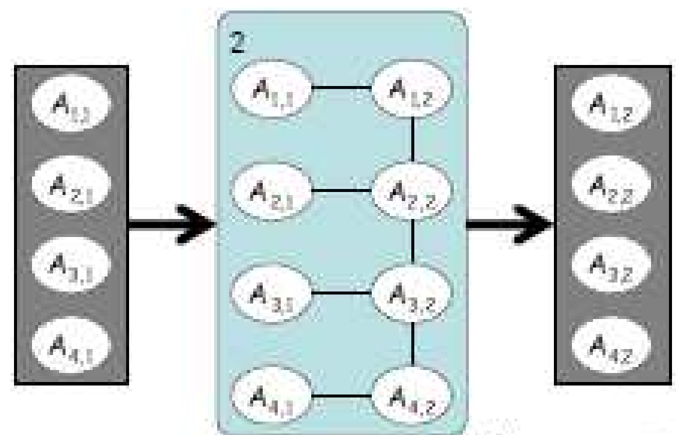
- Thm 11.4.5. Let Q be a set of beliefs st μ_{ij} is in the exp family Q_{ij}. Let M-project-distr$_{i,j}$ marginalize onto $S_{i,j}$ and then project onto Q_{ij}. Then Q is a stationary point of EP-optimize iff there exist auxiliary beliefs δ such that

$$\begin{aligned}
\delta_{i \to j} &= \frac{\text{M-project-distr}_{i,j}(\beta_i)}{\delta_{j \to i}} \\
\beta_i &\propto \psi_i \cdot \prod_{j \in \mathrm{Nb}_i} \delta_{j \to i} \\
\mu_{i,j} &\propto \delta_{j \to i} \cdot \delta_{i \to j}.
\end{aligned}$$

Example







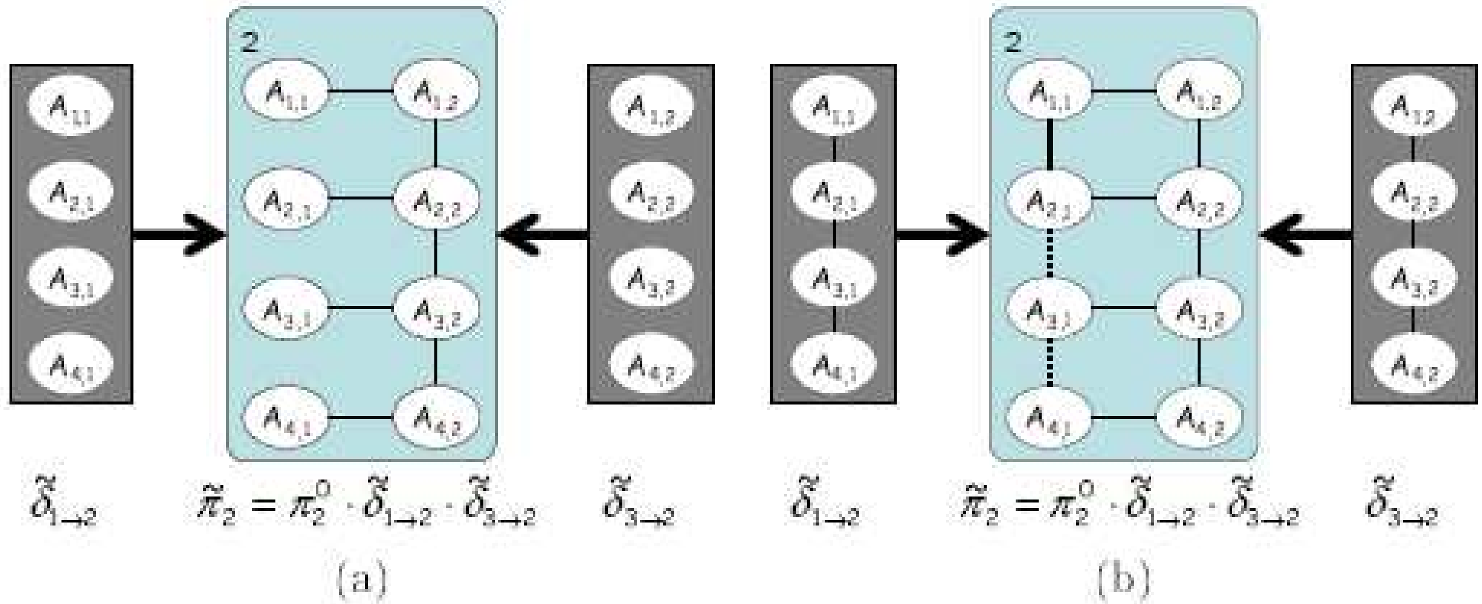
$\tilde{\delta}_{1\to2}$ π_2^0 $\tilde{\delta}_{2\to3} = \rho(\sum \pi_2^0 \cdot \tilde{\delta}_{1\to2})$

Cluster graph

Fully factored Q_{ij}

Structured messages

- The Q distribution (onto which we project) can be any structure that makes computing marginals efficient, eg a chain or clique tree.



Summary so far

- Let us summarize the BP and EP methods, and then introduce a new class of variational methods

BP on Cluster graphs

- In CGBP, we made 2 approximations
- 1. Optimize a bound on D(Q||P)

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

Thm 11.1.2

- 2. Use pseudo marginals β_I, μ_{ij} and thus approximated the entropy H(Q) and hence used the approximate bound

$$
\tilde{F}(\tilde{P}, Q) \stackrel{\text{def}}{=} \sum_i E_{C_i \sim \beta_i} \ln \psi_i + \sum_i H_{\beta_i}(C_i) - \sum_{<ij>} H_{\mu_{i,j}}(S_{i,j})
$$

- We then optimize the approximate bound subject to local consistency constraints over some cluster graph

CGBP objective

$$
\begin{array}{ll}
\textbf{Find} & Q = \{\beta_i : i \in \mathcal{V}_\mathcal{T}\} \cup \{\mu_{i,j} : (i\ j) \in \mathcal{E}_\mathcal{T}\} \\
\textbf{that maximize} & \tilde{F}[\tilde{P}_\Phi, Q]
\end{array}
$$

subject to

$$
\begin{array}{rcll}
\mu_{i,j}[s_{i,j}] & = & \displaystyle\sum_{\boldsymbol{C}_i - \boldsymbol{S}_{i,j}} \beta_i[c_i] & \\
& & \forall (i\ j) \in \mathcal{E}_\mathcal{T}, \forall s_{i,j} \in Val(\boldsymbol{S}_{i,j}) & \\
\displaystyle\sum_{\boldsymbol{c}_i} \beta_i[c_i] & = & 1 & \forall i \in \mathcal{V}_\mathcal{T} \\
\beta_i[c_i] & \geq & 0 & \forall i \in \mathcal{V}_\mathcal{T}, c_i \in Val(C_i)
\end{array}
$$

If the cluster graph is a cluster tree, this is exact

EP

- In EP, we make the same 2 approximations as in CGBP, but we also relax the local consistency constraint so that now cliques only have to agree on their expected sufficient statistics, not on their distributions

Find $\quad Q$

that maximize $\quad \tilde{F}[\tilde{P}_\Phi, Q]$

subject to

$$
\begin{aligned}
E_{\boldsymbol{S}_{i,j} \sim \mu_{i,j}}[\tau_{i,j}] &= E_{\boldsymbol{S}_{i,j} \sim \beta_j}[\tau_{i,j}] \quad \forall (i\text{–}j) \in \mathcal{E}_{\mathcal{T}} && (11.41)\\
\sum_{\boldsymbol{c}_i} \beta_i[\boldsymbol{c}_i] &= 1 \quad \forall i \in \mathcal{V}_{\mathcal{T}} && (11.42)\\
\sum_{\boldsymbol{s}_{i,j}} \mu_{i,j}[\boldsymbol{s}_{i,j}] &= 1 \quad \forall (i\text{–}j) \in \mathcal{E}_{\mathcal{T}} && (11.43)\\
\beta_i[\boldsymbol{c}_i] &\geq 0 \quad \forall i \in \mathcal{V}_{\mathcal{T}}, \boldsymbol{c}_i \in Val(\boldsymbol{C}_i) && (11.44)
\end{aligned}
$$

Even if the CG is a tree, this is no longer exact (in general)

Variational methods

- The problems with BP and EP are
  - They do not monotonically increase a lower bound on ln Z
  - They may not converge (except convex BP)
- Let us now require Q to be a coherent probability distribution (of tractable form). Hence we can now compute the exact entropy and optimize the exact objective

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

- This always increases the lower bound and will always converge

Mean field approximation

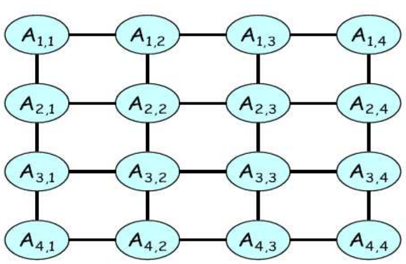
- Let us assume the approximate posterior is fully factorized

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

- Then the objective (negative free energy) is

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

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



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

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

$$\langle f(x_h) \rangle \stackrel{\text{def}}{=} \sum_{x_h} \left[\prod_{i \in h} Q_i(x_i) \right] f(x_h)$$

$$\langle f(x_h) \rangle_{j,k} \stackrel{\text{def}}{=} \sum_{x_h \backslash x_j} \left[\prod_{i \in h, i \neq j} Q_i(x_i) \right] f(x_h | x_j = k)$$

$$\langle f(x_h) \rangle = \sum_k Q_j(x_j = k) \langle f(x_h) \rangle_{j,k}$$

$$\ln p(x_v) \geq \sum_c \langle \ln \phi_c(x_c) \rangle + \sum_i H(Q_i)$$

$$= \sum_k Q_j(k) \sum_c \langle \ln \phi_c(x_c) \rangle_{j,k} + H(Q_j) + \sum_{i \neq j} H(Q_i)$$

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

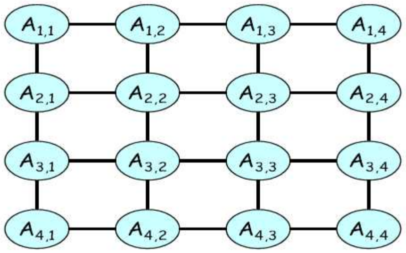
Mean field equations

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

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

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

- 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 $Q(\theta)$
- This gives us the benefits of being Bayesian for the same computational speed as EM

VB for univariate Gaussian

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

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

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

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

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

Gaussian

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

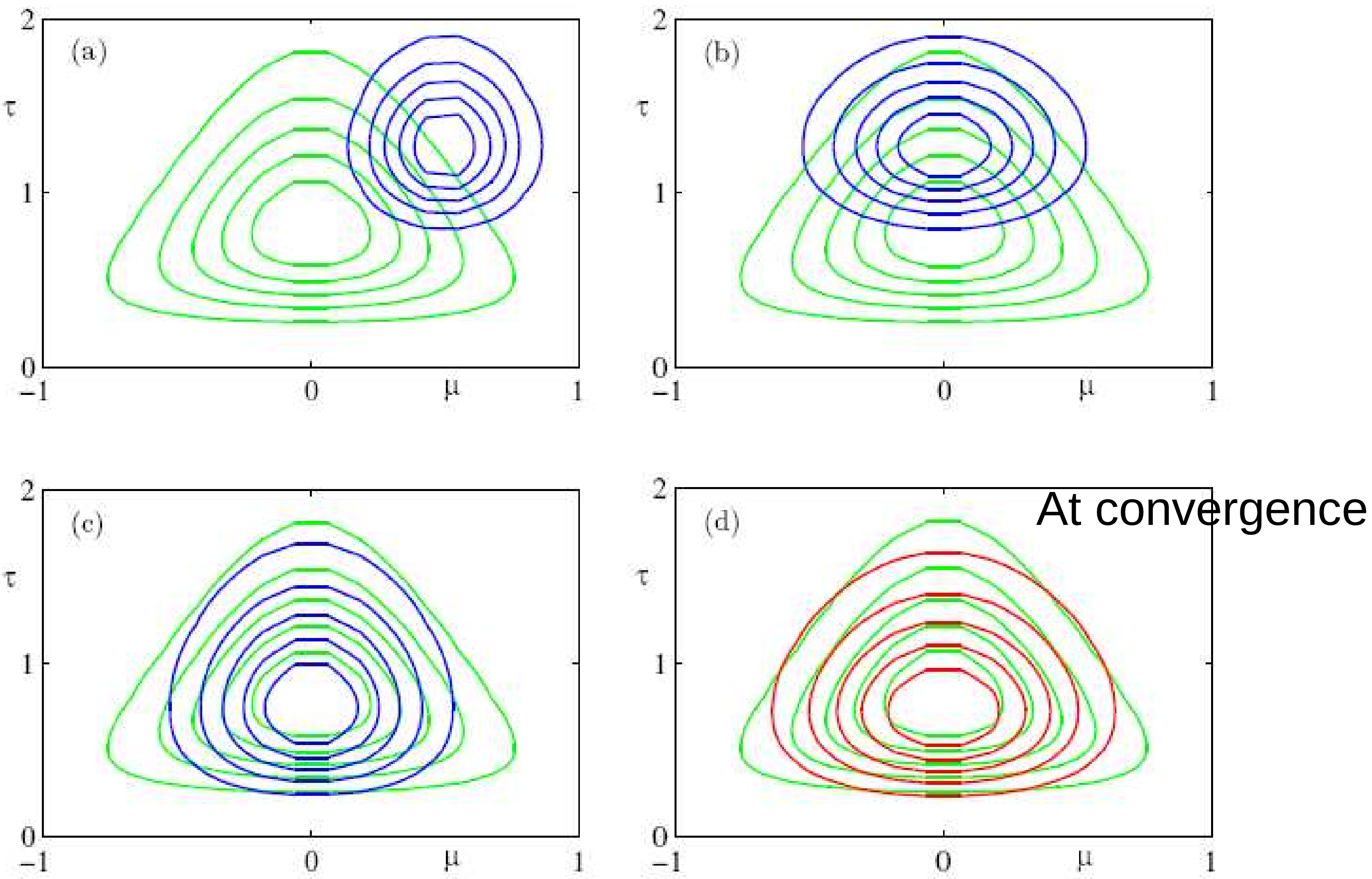
$$\begin{aligned} \ln q_\tau^\star(\tau) &= \mathbb{E}_\mu\left[\ln p(\mathcal{D}|\mu,\tau) + \ln p(\mu|\tau)\right] + \ln p(\tau) + \text{const.} \\ &= (a_0-1)\ln\tau - b_0\tau + \frac{N+1}{2}\ln\tau \\ &\quad -\frac{\tau}{2}\mathbb{E}_\mu\left[\sum_{n=1}^{N}(x_n-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right]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

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



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

VB for mixtures of Gaussians

Inference

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

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

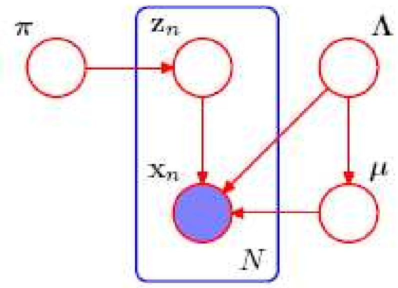
$$\ln q^\star(\mathbf{Z}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \ln \rho_{nk} + \text{const.}$$

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

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

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

Model



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

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

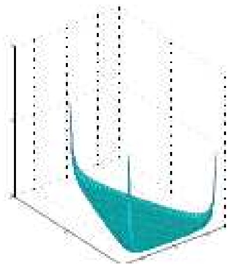
$$p(\mathbf{Z}|\boldsymbol{\pi}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k^{z_{nk}}.$$

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

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

Automatic model selection

- Recall $\pi \sim \mathrm{Dir}(\alpha)$. If $\alpha << 1$, we prefers skewed π and hence sparse z.



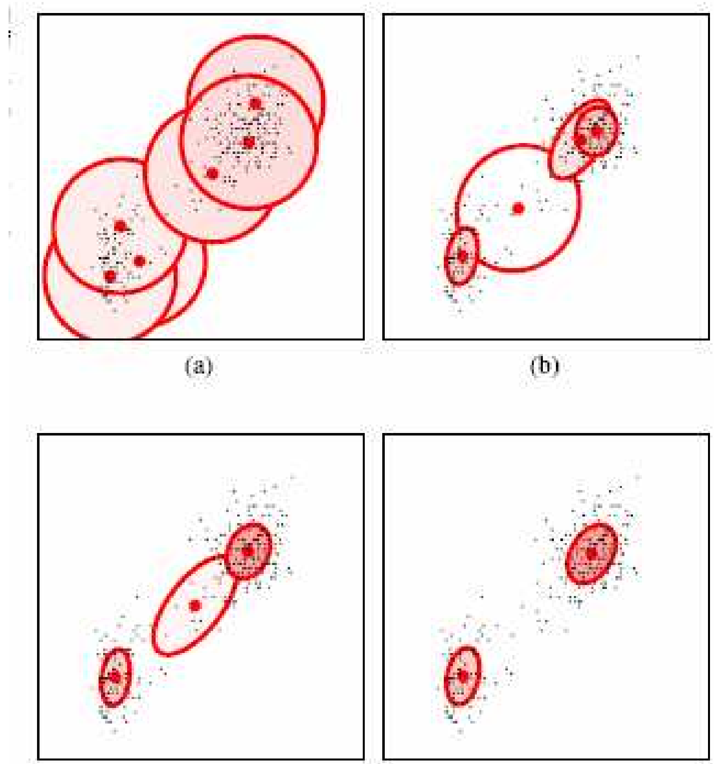
- MAP estimate from regular EM is

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

- Posterior mean estimate from VB is

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

Selecting K with one run of VB



Variational message passing

- Consider a DAG model

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

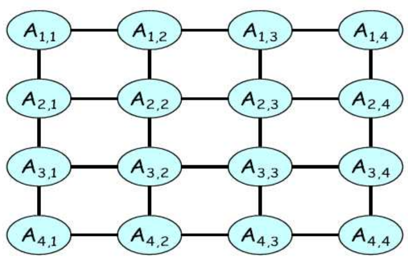
- The mean field equations are

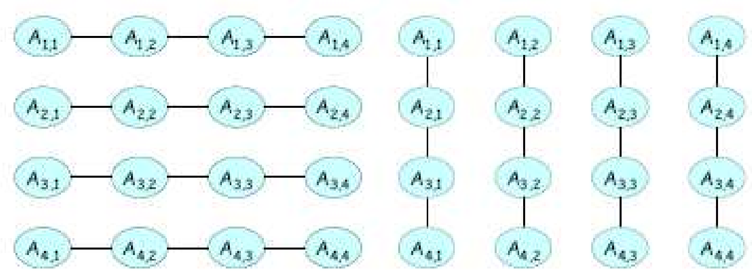
$$\ln q_j^\star(\mathbf{x}_j) = \mathbb{E}_{i \neq j}\left[\sum_i \ln p(\mathbf{x}_i | \mathrm{pa}_i)\right] + \mathrm{const.}$$

- 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





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

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:*

$$\psi_j(c_j) \propto \exp\left\{ E_Q\left[\ln \tilde{P}_\Phi \mid c_j\right] - \sum_{k \neq j} E_Q[\ln \psi_k \mid c_j] \right\}. \tag{11.59}$$