· Exact Inference as Optimization

Assume we have a factored distribution

$$P_{\Phi}(\mathcal{X}) = \frac{1}{Z} \prod_{\phi \in \Phi} \phi(U_{\phi})$$

In exact inference, we find a set of calibrated beliefs that represent $P_{\Phi}(\mathcal{X})$

That is, we find beliefs that match the distribution represented by given set of initial potentials

We can view exact inference as searching over the set of distribution Q that are representable by the cluster tree to find a distribution Q^* that matches P_{Φ}

↓

Searching for a calibrated distribution that is as close as possible to P_{Φ}

the KL-divergence (relative-divergence)

$$D(P_1 \| P_2) = \int_x p_1(x) \log \frac{P_1(x)}{P_2(x)} \, dx$$

Search for a distribution Q that minimizes $D(Q \| P_{\Phi})$

Suppose we are given a clique tree structure $\mathcal{T}$ for P_{Φ}
(i.e. $\mathcal{T}$ satisfies the running intersection property and family preserving property)
and a set of beliefs

$$Q = \{\beta_i \mid i \in V_{\mathcal{T}}\} \cup \{\mu_{ij} \mid (i-j) \in \mathcal{E}_{\mathcal{T}}\}$$

the set of beliefs in $\mathcal{T}$ defines a distribution Q

$$Q(\mathcal{X}) = \frac{\prod_{i \in V_{\mathcal{T}}} \beta_i}{\prod_{(i-j) \in \mathcal{E}_{\mathcal{T}}} \mu_{ij}}$$

if Q is a set of calibrated beliefs for $\mathcal{T}$, then $\begin{cases} \beta_i(C_i) = Q(C_i) \\ \mu_{ij}(S_{ij}) = Q(S_{ij}) \end{cases}$

find $Q = \{\beta_i \mid i \in V_{\mathcal{T}}\} \cup \{\mu_{ij} : (i-j) \in \mathcal{E}_{\mathcal{T}}\}$

minimize $D(Q \| P_{\Phi})$

s.t. $\mu_{ij}[S_{ij}] = \sum_{C_i - S_{ij}} \beta_i(C_i) \qquad \forall (i-j) \in \mathcal{E}_{\mathcal{T}} \quad \forall s_{ij} \in Val(S_{ij})$

$\sum_{C_i} \beta_i(C_i) = 1 \qquad \forall i \in V_{\mathcal{T}}$

· The Energy Functional

The objective $D(Q \| P_\Phi)$ is unwieldy for direct optimization, but we can rewrite the relative entropy in a simpler form

$$D(Q\|P_\Phi) = \int_x Q(x) \cdot \ln \frac{Q(x)}{P_\Phi(x)} \, dx = \ln Z - F[\tilde{P}_\Phi, Q]$$

where $F[\tilde{P}_\Phi, Q]$ is the energy functional

$$F[\tilde{P}_\Phi, Q] = E_Q[\ln \tilde{P}(x)] + H_Q(x) = \sum_{\phi \in \Phi} E_Q[\ln \phi] + H_Q(x)$$

and $H_Q(x) = -E_{x\sim Q}[\ln Q(x)] = -\int_x Q(x) \cdot \ln Q(x) \, dx$ is the entropy of Q

$$\begin{aligned} D(Q\|P_\Phi) &= \int_x Q(x) \cdot \ln \frac{Q(x)}{P_\Phi(x)} \, dx \\ &= \int_x Q(x) \cdot \ln Q(x) \, dx - \int_x Q(x) \cdot \ln P_\Phi(x) \, dx \\ &= E_{x\sim Q}[\ln Q(x)] - E_{x\sim Q}[\ln P_\Phi(x)] \end{aligned}$$

$P_\Phi(x) = \frac{1}{Z} \tilde{P}_\Phi(x) = \frac{1}{Z} \prod_{\phi \in \Phi} \phi(U_\phi)$ U_ϕ is the projection of x on $Scope[\phi]$

$$\begin{aligned} &= E_{x\sim Q}[\ln Q(x)] - E_{x\sim Q}[\sum_\phi \ln \phi(U_\phi) - \ln Z] \\ &= -H_Q(x) - \sum_{\phi \in \Phi} E_{x\sim Q}[\ln \phi(U_\phi)] + \ln Z \\ &= -F[\tilde{P}_\Phi, Q] + \ln Z \end{aligned}$$

Since Z is the normalizing constant, doesn't depend on Q.

$\therefore$ minimize KL-divergence $\rightarrow$ maximize $F[\tilde{P}_\Phi, Q]$

$$F[\tilde{P}_\Phi, Q] = \underbrace{-E_{x\sim Q}[\ln Q(x)]}_{\text{entropy term}} + \underbrace{\sum_{\phi \in \Phi} E_{x\sim Q}[\ln \phi(U_\phi)]}_{\text{energy term}}$$

· Optimizing the Energy Functional

$$\begin{aligned} D(Q\|P_\Phi) = \int_x Q(x) \cdot \ln \frac{Q(x)}{P_\Phi(x)} \, dx &= -H_Q(x) - \sum_{\phi \in \Phi} E_{x\sim Q}[\ln \phi(U_\phi)] + \ln Z \\ &= -F[\tilde{P}_\Phi, Q] + \ln Z \end{aligned}$$

$D(Q\|P_\Phi) \geq 0$ $\ln Z \geq F[\tilde{P}_\Phi, Q]$

the energy functional is a lower bound of the log of the partition function

Variational method: want to solve a problem by introducing new variational parameters that increase the degrees of freedom over which we optimize. Each choice of those parameters gives an approximate answer.

- Exact Inference as Optimization

Given a cluster tree $\mathcal{T}$ with a set of beliefs Q, define the factored energy functional

$$\tilde{F}[\tilde{P}_\Phi, Q] = \sum_{i \in V_\mathcal{T}} E_{C_i \sim \beta_i}[\ln \psi_i] + \sum_{i \in V_\mathcal{T}} H_{\beta_i}(C_i) - \sum_{(ij) \in \mathcal{E}_\mathcal{T}} H_{\mu_{ij}}(S_{ij})$$

where ψ_i is the initial potential assigned to C_i : $\psi_i = \prod_{\phi \in C_i} \phi$

$$\tilde{F}[\tilde{P}_\Phi, Q] = \sum_{i \in V_\mathcal{T}} E_{C_i \sim \beta_i}\left[\ln \prod_{\phi \in C_i} \phi\right] - \sum_{i \in V_\mathcal{T}} \int_{C_i} \beta_i(C_i) \ln \beta_i(C_i)\, dC_i + \sum_{(ij) \in \mathcal{E}_\mathcal{T}} \int_{S_{ij}} \mu_{ij}(S_{ij}) \ln \mu(S_{ij})\, dS_{ij}$$

1° $\displaystyle \sum_i E_{C_i \sim \beta_i}[\ln \psi_i] = \sum_i \int_{C_i} \beta_i(C_i) \ln \prod_{\phi \in C_i} \phi(U_\phi)\, dC_i$

$$= \sum_i \sum_{\phi \in C_i} \int_{C_i} Q(C_i) \ln \phi(U_\phi)\, dC_i$$

$$= \sum_{\phi \in \Phi} E_{C_i \sim Q}[\ln \phi]$$

2° $\displaystyle H_Q(\mathcal{X}) = -\int_x Q(x) \cdot \ln Q(x)\, dx = -\int_x Q(x) \cdot \ln \frac{\prod_{i \in V_\mathcal{T}} \beta_i(C_i)}{\prod_{ij \in \mathcal{E}_\mathcal{T}} \mu_{ij}(S_{ij})}\, dx$

$$= -\int_x Q(x) \cdot \sum_{i \in V_\mathcal{T}} \ln \beta_i(C_i)\, dx + \int_x Q(x) \sum_{ij \in \mathcal{E}_\mathcal{T}} \ln \mu_{ij}(S_{ij})\, dx$$

$$= -\sum_{i \in V_\mathcal{T}} \int_x Q(x) \ln \beta_i(C_i)\, dx + \sum_{ij \in \mathcal{E}_\mathcal{T}} \int_x Q(x) \ln \mu_{ij}(S_{ij})\, dx$$

$$= -\sum_{i \in V_\mathcal{T}} \int_{C_i} \int_{\mathcal{X}-C_i} Q(C_i, \mathcal{X}-C_i) \ln \beta_i(C_i)\, d(\mathcal{X}-C_i)\, dC_i + \sum_{ij \in \mathcal{E}_\mathcal{T}} \int_{S_{ij}} \int_{\mathcal{X}-S_{ij}} Q(S_{ij}, \mathcal{X}-S_{ij}) \ln \mu_{ij}(S_{ij})\, d(\mathcal{X}-S_{ij})\, dS_{ij}$$

$$= -\sum_{i \in V_\mathcal{T}} \int_{C_i} \beta_i(C_i) \ln \beta_i(C_i)\, dC_i + \sum_{(ij) \in \mathcal{E}_\mathcal{T}} \int_{S_{ij}} \mu_{ij}(S_{ij}) \ln \mu_{ij}(S_{ij})\, dS_{ij}$$

$$= \sum_{i \in V_\mathcal{T}} H_{\beta_i}(C_i) - \sum_{(ij) \in \mathcal{E}_\mathcal{T}} H_{\mu_{ij}}(S_{ij})$$

$$F[\tilde{P}_\Phi, Q] = \sum_{\phi \in \Phi} E_{\mathcal{X} \sim Q}[\ln \phi] + H_Q(\mathcal{X}) = \sum_i E_{C_i \sim \beta_i}[\ln \psi_i] + \sum_{i \in V_\mathcal{T}} H_{\beta_i}(C_i) - \sum_{(ij) \in \mathcal{E}_\mathcal{T}} H_{\mu_{ij}}(S_{ij}) = \tilde{F}[\tilde{P}_\Phi, Q]$$

Find $Q = \{\beta_i \mid i \in V_\mathcal{T}\} \cup \{\mu_{ij} \mid (ij) \in \mathcal{E}_\mathcal{T}\}$

max. $F[\tilde{P}_\Phi, Q] = \tilde{F}[\tilde{P}_\Phi, Q]$

s.t. $\mu_{ij}(s_{ij}) = \sum_{C_i - S_{ij}} \beta_i(C_i)$ $\quad \forall (ij) \in \mathcal{E}_\mathcal{T},\ \forall s_{ij} \in Val(S_{ij})$ $\quad \}$ marginal consistency

$\sum_{c_i} \beta_i(c_i) = 1$ $\quad \forall i \in V_\mathcal{T}$

$\beta_i(c_i) \geq 0$ $\quad \forall i \in V_\mathcal{T} \quad \forall c_i \in Val(C_i)$ $\quad \}$ joint probability

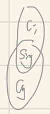
· Fixed-Point Characterization

If a clique tree $\mathcal{T}$ is an I-map of P_Φ, then there's a unique solution to

$$
\begin{aligned}
\text{Find} \quad & Q = \{\beta_i \mid i \in V_\mathcal{T}\} \cup \{\mu_{ij} : (ij) \in \mathcal{E}_\mathcal{T}\} \\
\text{min.} \quad & D(Q \| P_\Phi) = -F[\tilde{P}_\Phi, Q] + \ln(Z) = -\tilde{F}[\tilde{P}_\Phi, Q] \\
\text{s.t.} \quad & \mu_{ij}(S_{ij}) = \sum_{C_i - S_{ij}} \beta_i(C_i) \quad \forall (i\text{-}j) \in \mathcal{E}_\mathcal{T} \quad \forall s_{ij} \in Val(S_{ij})
\end{aligned}
$$

$$
\begin{aligned}
D(Q \| P_\Phi) &= \int_x Q(x) \ln \frac{Q(x)}{P_\Phi(x)} dx \\
&= \int_x Q(x) \ln Q(x)\, dx - \int_x Q(x) \ln P_\Phi(x)\, dx \\
&= -H_Q(\mathcal{X}) - E_{x\sim Q}[\ln P_\Phi(\mathcal{X})] \qquad P_\Phi(\mathcal{X}) = (\textstyle\prod_\phi \phi)(\mathcal{X}) \cdot \frac{1}{Z} = \frac{1}{Z}\tilde{P}_\Phi(\mathcal{X}) \\
&= -H_Q(\mathcal{X}) - E_{x\sim Q}[\ln (\textstyle\prod_\phi \phi)(x)] + \ln Z \\
&= -H_Q(\mathcal{X}) - \sum_{\phi\in\Phi} E_{x\sim Q}[\ln \phi(U_\phi)] + \ln Z \\
&= -F[\tilde{P}_\Phi, Q] + \ln Z \qquad \text{Energy functional: } F[\tilde{P}_\Phi, Q] = \sum_\phi E_Q[\ln \phi] + H_Q(\mathcal{X}) \\
&= -\sum_{i\in V_\mathcal{T}} E_{C_i\sim\beta_i}[\ln \psi_i] - \sum_{i\in V_\mathcal{T}} H_{\beta_i}(C_i) + \sum_{(ij)\in\mathcal{E}_\mathcal{T}} H_{\mu_{ij}}(S_{ij}) + \ln Z \\
&= -\tilde{F}[\tilde{P}_\Phi, Q] + \ln Z \qquad \text{Factored energy functional } \tilde{F}[\tilde{P}_\Phi, Q] = \sum_{i\in V_\mathcal{T}} E_{C_i\sim\beta_i}[\ln \psi_i] + \sum_{i\in V_\mathcal{T}} H_{\beta_i}(C_i) - \sum_{(ij)\in\mathcal{E}_\mathcal{T}} H_{\mu_{ij}}(S_{ij})
\end{aligned}
$$

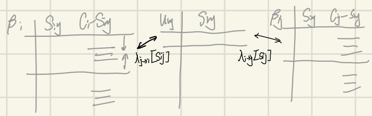
$$
\begin{aligned}
\text{Find} \quad & Q = \{\beta_i \mid i \in V_\mathcal{T}\} \cup \{\mu_{ij} \mid ij \in \mathcal{E}_\mathcal{T}\} \\
\text{min.} \quad & -\sum_{i\in V_\mathcal{T}} \int_{C_i} \beta_i(c_i) \ln \psi_i(c_i)\, dc_i + \sum_{i\in V_\mathcal{T}} \int_{C_i} \beta_i(c_i) \ln \beta_i(c_i)\, dc_i - \sum_{(ij)\in\mathcal{E}_\mathcal{T}} \int_{S_{ij}} \mu_{ij}(s_{ij}) \ln \mu_{ij}(s_{ij})\, ds_{ij} \\
\text{s.t.} \quad & \mu_{ij}[s_{ij}] = \Big(\sum_{C_i - S_{ij}} \beta_i\Big)(s_{ij}) \qquad \forall (ij) \in \mathcal{E}_\mathcal{T} \quad \forall s_{ij} \in Val(S_{ij}) \qquad \text{dual variable } \lambda_{j\to i}[s_{ij}] \\
& \sum_{c_i} \beta_i(c_i) = 1 \qquad \forall i \in V_\mathcal{T} \qquad \text{dual variable } \lambda_i \\
& \beta_i(c_i) \geq 0 \qquad \forall i \in V_\mathcal{T}, \quad c_i \in Val(C_i) \qquad \text{// implicitly satisfied}
\end{aligned}
$$



$$
\mathcal{L} = -\tilde{F}[\tilde{P}_\Phi, Q] + \sum_{i\in V_\mathcal{T}} \lambda_i \Big(\sum_{c_i} \beta_i(c_i) - 1\Big) + \sum_i \sum_{j\in N_{\beta_i}} \sum_{s_{ij}} \lambda_{j\to i}[s_{ij}] \cdot \Big(\mu_{ij}[s_{ij}] - \Big(\sum_{C_i - S_{ij}} \beta_i\Big)(s_{ij})\Big)
$$

$$
\frac{\partial \mathcal{L}}{\partial \beta_i(c_i)} = -\ln \psi_i(c_i) + \ln \beta_i(c_i) + 1 + \lambda_i - \sum_{j\in N_{\beta_i}} \lambda_{j\to i}[s_{ij}] := 0
$$

$$
\beta_i(c_i) = \exp(-1-\lambda_i) \cdot \psi_i(c_i) \cdot \prod_{j\in N_{\beta_i}} \exp(\lambda_{j\to i}[s_{ij}])
$$



$$
\frac{\partial \mathcal{L}}{\partial \mu_{ij}[s_{ij}]} = -\ln \mu_{ij}(s_{ij}) - 1 + \lambda_{j\to i}[s_{ij}] + \lambda_{i\to j}[s_{ij}] := 0
$$

$$
\mu_{ij}(s_{ij}) = \exp(-1) \cdot \exp(\lambda_{j\to i}[s_{ij}]) \cdot \exp(\lambda_{i\to j}[s_{ij}])
$$

let $\delta_{i\to j}[s_{ij}] = \exp\big(\lambda_{i\to j}[s_{ij}] - \frac{1}{2}\big)$

$$
\mu_{ij}(s_{ij}) = \delta_{i\to j}(s_{ij}) \cdot \delta_{j\to i}(s_{ij})
$$

$$
\beta_i(c_i) = \exp(-1-\lambda_i) \cdot \psi_i(c_i) \prod_{j\in N_{\beta_i}} \exp\big(\lambda_{j\to i}[s_{ij}] - \tfrac{1}{2}\big) \cdot \exp\Big(\frac{|N_{\beta_i}|}{2}\Big)
$$

$$\beta_i(C_i) = \exp\left(-1-\lambda_i + \frac{|NB_i|}{2}\right) \cdot \psi_i(C_i) \cdot \prod_{j \in NB_i} \delta_{j\to i}(S_{ij})$$

$$\delta_{i\to j}(S_{ij}) = \frac{\mu_{ij}(S_{ij})}{\delta_{j\to i}(S_{ij})} = \frac{\left(\sum_{C_i - S_{ij}} \beta_i\right)(S_{ij})}{\delta_{j\to i}(S_{ij})}$$

$$= \exp\left(-1-\lambda_i + \frac{|NB_i|}{2}\right) \cdot \left(\sum_{C_i - S_{ij}} \psi_i\right)(S_{ij}) \prod_{k \in \{NB_i - j\}} \delta_{k\to i}(S_{ij})$$

$$\delta_{i\to j}(S_{ij}) = C \cdot \left(\sum_{C_i - S_{ij}} \psi_i\right) \bowtie \prod_{k \in \{NB_i - j\}} \delta_{k\to i}(S_{ij})$$ belief propagation updating rule

A set of beliefs Q is a stationary point of the optimization problem iff

$$\exists \{ \delta_{i\to j}[S_{ij}] \mid (i,j) \in \mathcal{E}_T \}$$

s.t.

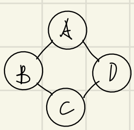
$$\delta_{i\to j} \propto \sum_{C_i - S_{ij}} \psi_i \left(\prod_{k \in \{NB_i - j\}} \delta_{k\to i} \right)$$

moreover, we have

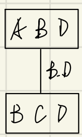
$$\beta_i \propto \psi_i \left(\prod_{j \in NB_i} \delta_{j\to i} \right)$$

$$\mu_{ij} = \delta_{i\to j} \cdot \delta_{j\to i}$$

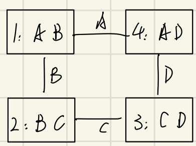
- Propagation-Based Approximation : Example



Markov Network



A clique tree



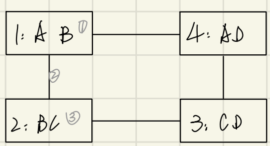
A cluster graph

the cluster graph contains loops (loopy), we can still apply belief-update propagation (nothing in the algorithm relies on the fact that it's a tree)

1° $\delta_{1\to2}(B) = \sum_A \psi_1(A,B)$

$\beta_2(B,C) = \beta_2(B,C) \cdot \delta_{1\to2}(B) / \mu_{12}(B) = \psi_2(B,C) \cdot \sum_A \psi_1(A,B)$

$\mu_{12}(B) = \delta_{1\to2}(B) = \sum_A \psi_1(A,B)$



① $\beta_1(A,B) = \psi_1(A,B)$

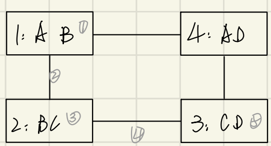
② $\mu_{12}(B) = \delta_{1\to2}(B) = \sum_A \beta_1 = \sum_A \psi_1(A,B)$

③ $\beta_2(B,C) = \psi_2(B,C) \cdot \sum_A \psi_1(A,B)$

2° $\delta_{2\to3}(C) = \sum_B \beta_2 = \sum_B \psi_2(B,C) \cdot \sum_A \psi_1(A,B)$

$\beta_3(C,D) = \beta_3(C,D) \cdot \delta_{2\to3}(C) / \mu_{23}(C) = \psi_3(C,D) \sum_B \psi_2(B,C) \sum_A \psi_1(A,B)$

$\mu_{23}(C) = \delta_{2\to3}(C) = \sum_B \psi_2(B,C) \cdot \sum_A \psi_1(A,B)$



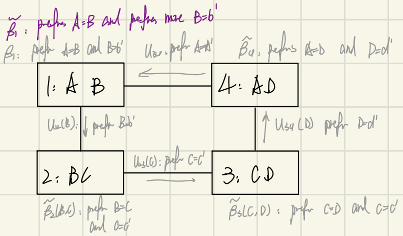
① $\beta_1(A,B) = \psi_1(A,B)$

② $\mu_{12}(B) = \delta_{1\to2}(B) = \sum_A \beta_1 = \sum_A \psi_1(A,B)$

③ $\beta_2(B,C) = \psi_2(B,C) \cdot \sum_A \psi_1(A,B)$

④ $\mu_{23}(C) = \delta_{2\to3}(C) = \sum_B \psi_2(B,C) \cdot \sum_A \psi_1(A,B)$

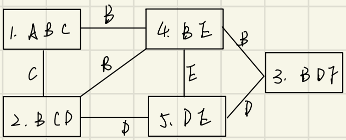
if all clusters favor consensus joint assignments (i.e. $\beta_1(a^0 b^0)$ and $\beta_1(a^1 b^1) \gg \beta_1(a^1, b^0)$ and $\beta_1(a^0, b^1)$)

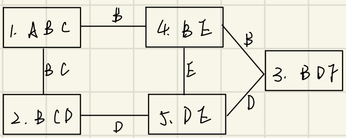


the process may not converge

for cluster

· Cluster-Graph Belief Propagation





A (loopy) cluster graph $\mathcal{U}$ satisfies the running intersection property if
whenever there is a variable such that $X \in C_i$ and $X \in C_j$,
then there is a single path between C_i and C_j for which $X \in S_e$ $\forall e$ in the path

running intersection property:
1° $\exists$ a path: information about X flow between all clusters that contain it
so that in a calibrated cluster graph, all cluster must agree about the marginal of X
2° only one path: prevents information about X from cycling endlessly in a loop

A cluster graph is calibrated if
$\forall (i\text{-}j) \in \mathcal{E}_{\mathcal{U}}$, $\sum_{C_i - S_{ij}} \beta_i = \sum_{C_j - S_{ij}} \beta_j$ (two clusters agrees on marginal of variables in S_{ij})
not necessarily all vars 2 cliques have in common

def initialize_cluster_graph($\mathcal{U}$):
for each cluster C_i:
$\beta_i = \prod_{\phi \in C_i} \phi$
for each $(i\text{-}j) \in \mathcal{E}_{\mathcal{U}}$
$\delta_{i \to j} = 1$ $\delta_{j \to i} = 1$

def sum_product_message (i: sending clique , j: receiving clique):
$\psi_i(C_i) = \psi_i \cdot \prod_{k \in \mathcal{N}_i - \{j\}} \delta_{k \to i}$
$\tau(S_{ij}) = \sum_{C_i - S_{ij}} \psi(C_i)$
return $\tau(S_{ij})$

def cluster_graph_sum_product_calibrate(Φ: set of factors, $\mathcal{U}$: cluster Graph):
initialize_cluster_graph($\mathcal{U}$)
while graph is not calibrated
select $(i,j) \in \mathcal{E}_{\mathcal{U}}$
$\delta_{i \to j}(S_{ij})$ = sum_product_message(i,j)
for each clique i
$\beta_i \leftarrow \psi_i \cdot \prod_{k \in \mathcal{N}_i} \delta_{k \to i}$
return $\{\beta_i\}$

other than the fact that this algorithm is applied to graphs rather than trees,
this algorithm is identical to sum-product calibration of clique trees (initialize all messages to 1)

can use belief-update messages to define Belief-update calibration for cluster graphs

```
def initialize_cluster_graph():
    for each cluster Ci:
        βi = ∏ {φ | φ ∈ Ci}
    for each edge (i-j) ∈ Eu
        μi,j = 1
```

```
def belief_update_message(i: sending clique, j: receiving clique):
    σi→j = Σ_{Ci - Si,j} βi
    βj = βj · σi→j / μi,j
    μi,j = σi→j
```

```
def cluster_graph_belief_update_calibrate(Φ: set of factors, U: cluster graph over Φ):
    initialize_cluster_graph()
    while graph is not calibrated:
        select (i-j) ∈ Eu
        belief_update_message(i, j)
    return {βi} {μi,j}
```

· Properties of Cluster-Graph Belief propagation: Reparameterization

Let U be a generalized cluster graph over a set of factors Φ,
Consider the set of beliefs $\{\beta_i\}$ and sepsets $\{\mu_{i,j}\}$ at any iteration of CG-BP-Calibrate

$$\tilde{P}_\Phi(\mathcal{X}) = \frac{\prod_{i \in V_U} \beta_i(C_i)}{\prod_{(i-j) \in E_U} \mu_{i,j}(S_{i,j})}$$

where $\tilde{P}_\Phi(\mathcal{X}) = \prod_{\phi \in \Phi} \phi$ is the unnormalized distribution defined by Φ

At each iteration: $\beta_i = \psi_i \prod_{j \in Nb_i} \delta_{j \to i}$ $\quad\quad \mu_{i,j}(S_{i,j}) = \delta_{j \to i} \cdot \delta_{i \to j}$

$$\frac{\prod_{i \in V_U} \beta_i(C_i)}{\prod_{(i-j) \in E_U} \mu_{i,j}(S_{i,j})} = \frac{\prod_{i \in V_U} \psi_i(C_i) \prod_{j \in Nb(i)} \delta_{j \to i}(S_{i,j})}{\prod_{(i-j) \in E_U} \delta_{i \to j}(S_{i,j})\, \delta_{j \to i}(S_{i,j})}$$

$$= \prod_{i \in V_U} \psi_i(C_i)$$

$$= \tilde{P}_\Phi(\mathcal{X})$$

cluster-graph belief propagation preserves all the information about the original distribution
it does not "dilute" the original factors by performing propagation along loops.
This process kind of tries to represent the original factors in a more useful form

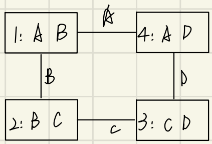
· Properties of Cluster-Graph Belief propagation: Tree Consistency

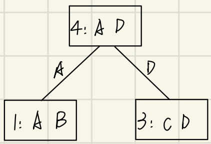
In a calibrated cluster Tree, the belief over a cluster is the marginal of the distribution

To characterize the beliefs we get by calibrating a cluster graph,

we can use the cluster tree invariant property applied to subtrees of a cluster graph

A subtree T of U is a subset of clusters and edges that together form a tree that satisfies the running intersection property





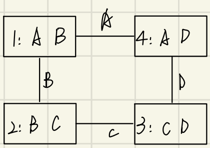
removing edges from a cluster graph may result in violating the running intersection property

Once we select a tree T, we can think of it as defining a distribution $P_T(X) = \frac{\prod_{i \in V_T} \beta_i(C_i)}{\prod_{(i-j) \in E_T} \mu_{i,j}(S_{i,j})}$

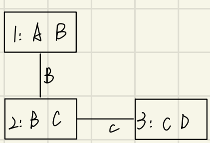
if the cluster graph is calibrated. so is the sub cluster tree, which satisfies the running intersection property

$\therefore \ \beta_i(C_i) = P_T(C_i)$

Tree consistency: The beliefs over C_i in the tree are the marginals of P_T



cluster graph U



cluster tree T (also a (calibrated) cluster graph)

$P_\Phi(A,B,C,D) = \frac{\beta_1(A,B)\cdot\beta_2(B,C)\cdot\beta_3(C,D)\cdot\beta_4(D,A)}{\mu_{12}(B)\cdot\mu_{23}(C)\cdot\mu_{34}(D)\cdot\mu_{41}(A)}$

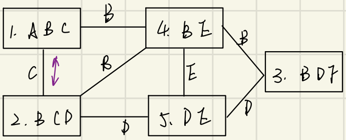
$P_T(A,B,C,D) = P_\Phi(A,B,C,D)\cdot\frac{\mu_{34}(D)\cdot\mu_{41}(A)}{\beta_4(D,A)}$

· Constructing Cluster Graphs

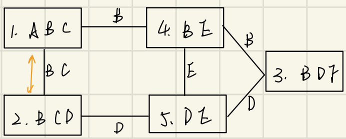
In the case of clique trees, different clique trees of a graphical model give the same answer

In the case of cluster graphs, different cluster graphs can lead to different answers

We have to consider the accuracy-cost trade-offs



messages about C



message about B C

Assume $C_1(A,B,C)$ strongly prefer $B=C$

U_2: the correlation is directly conveyed from C_1 to C_2

U_1: marginal on C is conveyed on (1,2), marginal on B is conveyed on (1-4-3)

the strong dependency is lost

if marginal of C in $C_1(A,B,C)$ is uniform, then $\delta_{1\to4}(B)$ is uniform,

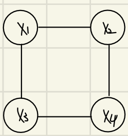
in U_1, B and C seems to be two independent uniform to C_2

eg. pairwise Markov Network

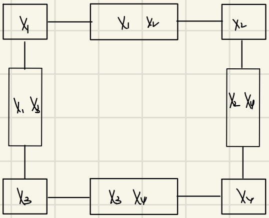
univariate potential $\phi_i(X_i)$ for each variable X_i

pairwise potential $\phi_{ij}(X_i, X_j)$ over some pairs of variables

(any distribution can be reformulated as a pairwise Markov network with transformation of variables)



pairwise Markov network



cluster graph

Variational Analysis

for some cluster graph $\mathcal{U}$ and a distribution P, define

$$Q_P = \{ P(C_i) \mid i \in \mathcal{V}_\mathcal{U} \} \cup \{ P(S_{ij}) \mid \langle i-j \rangle \in \mathcal{E}_\mathcal{U} \}$$

define the marginal polytope

$$\mathrm{Marg}[\mathcal{U}] = \{ Q_P : P \text{ is a distribution over } \mathcal{X} \}$$

The marginal polytope is the set of all cluster and sepset beliefs that can be obtained by marginalizing an actual P

(there are calibrated cluster beliefs that do not represent the marginals of any single coherent distribution over $\mathcal{X}$)

The marginal polytope has exponentially many facet, cannot optimize over $\mathrm{Marg}[\mathcal{U}]$

instead, we optimize the

$$\mathrm{Local}[\mathcal{U}] = \left\{ \begin{array}{l} \{\beta_i \mid i \in \mathcal{V}_\mathcal{U}\} \cup \\ \{\mu_{ij} \mid ij \in \mathcal{E}_\mathcal{U}\} \end{array} \middle| \begin{array}{lll} \mu_{ij}[s_{ij}] = \sum_{C_i \sim s_{ij}} \beta_i(c_i) & \forall \langle ij \rangle \in \mathcal{E}_\mathcal{U}, & s_{ij} \in Val(S_{ij}) \\ \sum_{c_i} \beta_i(c_i) = 1 & \forall i \in \mathcal{V}_\mathcal{U} & \\ \beta_i(c_i) \geq 0 & \forall i \in \mathcal{V}_\mathcal{U} & c_i \in Val(C_i) \end{array} \right\}$$

$$\begin{aligned} D(Q \| P_\Phi) &= \int_x Q(x) \cdot \ln \frac{Q(x)}{P_\Phi(x)} dx \\ &= \int_x Q(x) \cdot \ln Q(x)\, dx - \int_x Q(x) \cdot \ln P_\Phi(x)\, dx \\ &= -H_Q(\mathcal{X}) - E_{x \sim Q}\left[\ln \prod_\phi \phi(x)\right] + \ln Z \\ &= -H_Q(\mathcal{X}) - \sum_\phi E_{x \sim Q}[\ln \phi(x)] + \ln Z \\ &= -F[\tilde{P}_\Phi, Q] + \ln Z \end{aligned}$$

$$-\tilde{F}[\tilde{P}_\Phi, Q] + \ln Z = D(Q \| P_\Phi) \geq 0$$

$$\tilde{F}[\tilde{P}_\Phi, Q] \leq \ln Z$$

unlike for clique tree, $\tilde{F}[\tilde{P}_\Phi, Q]$ is no longer a reformulation of the energy function, but an approximation of the energy function

Find	Q
max.	$\tilde{F}[\tilde{P}_\Phi, Q]$
s.t.	$Q \in \mathrm{Local}[\mathcal{U}]$

Same optimization problem as for clique tree calibration,

but with 2 approximations (relaxations)

1° Use the factored energy as an approximation of the true energy function

2° optimize over space of pseudo-marginals instead of the space of all coherent distribution

A set of beliefs Q is a stationary point of the optimization problem iff

$\forall \langle ij \rangle \in \mathcal{E}_\mathcal{U}$, there are auxiliary factors $\delta_{i \to j}(S_{ij})$ and $\delta_{j \to i}(S_{ij})$ s.t. $\delta_{i \to j} \propto \sum_{C_i - S_{ij}} \psi_i \cdot \prod_{k \in (Nb_i - \{j\})} \delta_{k \to i}$

and we have

$$\beta_i \propto \psi_i \prod_{j \in Nb(i)} \delta_{j \to i}$$

$$\mu_{ij} = \delta_{i \to j} \cdot \delta_{j \to i}$$

· Structured Variational Approximations

The structured variational approach aims to optimize the energy functional over a family $\mathcal{Q}$ of coherent distribution Q. This family is chosen to be computationally tractable. hence it's generally not sufficiently expressive to capture all of the information in $\tilde{P}_{\Phi}$

Find. $Q \in \mathcal{Q}$

max. $F[\tilde{P}_{\Phi}, Q]$

$\mathcal{Q}$ is a given family of distribution.

$F[\tilde{P}_{\Phi}, Q]$ is the exact energy functional, thus maximizing $F[\,] \leftrightarrow$ minimizing $D(Q \| P_{\Phi})$

choose simple Q:

∵ Q can be described by a BN or MN with small tree-width, more efficient inference

∴ poor approximation of P_{Φ}

· Structured Variational Approximation: Mean Field Approximation

The mean field approximation finds the distribution Q that's closest to P_{Φ} in terms of $D(Q \| P_{\Phi})$ within the class of distributions representable as a product of independent marginals

$$Q(\mathcal{X}) = \prod_i Q(X_i)$$

the approximation of P_{Φ} as a fully factored distribution is likely to lose a lot of information in P_{Φ}

The energy functional $F[\tilde{P}_{\Phi}, Q] = H_Q(\mathcal{X}) + \sum_{\phi \in \Phi} E_{\mathcal{X} \sim Q}[\ln \phi(U_\phi)]$

$$E_{\mathcal{X} \sim Q}[\ln \phi(U_\phi)] = E_{U_\phi \sim Q}[\ln \phi(U_\phi)] = \int_{u_\phi \in Val(U_\phi)} Q(u_\phi) \cdot \ln \phi(u_\phi)\, du_\phi$$

$$= \int_{u_\phi \in Val(U_\phi)} \prod_{X_i \in U_\phi} Q(x_i) \ln \phi(u_\phi)\, du_\phi$$

$$H_Q(\mathcal{X}) = -\int_{x \in Val(\mathcal{X})} Q(x) \ln Q(x)\, dx = -\int_{x \in Val(\mathcal{X})} \prod_i Q(x_i) \cdot \ln \prod_i Q(x_i)\, dx$$

$$= -\sum_i \int_{x \in Val(\mathcal{X})} \prod_j Q(x_j) \ln Q(x_i)\, dx$$

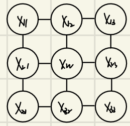
$$= -\sum_i \int_{x_1} Q(x_1) \cdots \int_{x_1} Q(x_i) \ln Q(x_i)\, dx_i \cdots dx$$ // $Q(x_i)$ is a marginal and integrates to 1

$$= -\sum_i \int_{x_i} Q(x_i) \ln Q(x_i)\, dx_i$$

$$= \sum_i H_Q(X_i)$$

if $Q(\mathcal{X}) = \prod_i Q(X_i)$, then $H_Q(\mathcal{X}) = \sum_i H_Q(X_i)$

eg. 4x4 grid Markov network



$$\tilde{F}[\tilde{P}_\Phi, Q] = H_Q(\mathcal{X}) + \sum_{\phi\in\Phi} E_{\mathcal{X}\sim Q}[\ln\phi(U_\phi)]$$

$$= \sum_{i\in\{1,2,3\}} \sum_{j\in\{1,2,3\}} E_Q[\ln\phi(X_{ij}, X_{i+1,j})]$$ // vertical potentials

$$+ \sum_{i\in\{1,2,3\}} \sum_{j\in\{1,2,3\}} E_Q[\ln\phi(X_{ij}, X_{i,j+1})]$$ // horizontal potentials

$$+ \sum_{i\in\{1:3\}} \sum_{j\in\{1:3\}} H_Q(X_{ij})$$ // node potentials

Find $\{Q(X_i)\}$

min. $-\tilde{F}[\tilde{P}_\Phi, Q] = \sum_i \int_{x_i} Q(x_i)\cdot\ln Q(x_i)\,dx_i - \sum_{\phi\in\Phi} \int_{u_\phi\in Val(U_\phi)} \prod_{x_i\in u_\phi} Q(x_i)\ln\phi(u_\phi)\,du_\phi$

s.t. $\sum_{x_i} Q(x_i) = 1 \quad \forall i$

$D(Q\|P_\Phi) = \int_x Q(x)\cdot\ln\frac{Q(x)}{P_\Phi(x)}\,dx$ is convex in $Q(x)$ $\forall x\in Val(\mathcal{X})$

$Q(x) = \prod_i Q(x_i)$ is jointly convex in $Q(x_i)$ and increasing $\forall x_i$

$D(Q\|P_\Phi) = -\tilde{F}[\tilde{P}_\Phi, Q] + \ln Z$ $\therefore$ $-\tilde{F}[\tilde{P}_\Phi, Q]$ is convex

$$\mathcal{L} = \sum_i \int_{x_i} Q(x_i)\cdot\ln Q(x_i)\,dx_i - \sum_{\phi\in\Phi}\int_{u_\phi} \prod_{x_i\in u_\phi} Q(x_i)\ln\phi(U_\phi)\,du_\phi + \sum_i \lambda_i\cdot\Big(\sum_{x_i} Q(x_i) - 1\Big)$$

$\phi \mid x_i \; \{x_{ij}\}$

$$\frac{\partial\mathcal{L}}{\partial Q(x_i)} = 1 + \ln Q(x_i) - \sum_{\phi\in\Phi}\int_{\substack{u_\phi\in Val(U_\phi)\\ X_i = x_i}} \prod_{\substack{x_j\in u_\phi\\ j\neq i}} Q(x_j)\ln\phi(u_\phi)\,du_\phi + \lambda$$

$$= 1 + \ln Q(x_i) - \sum_{\phi\in\Phi} E_{\mathcal{X}\sim Q}[\ln\phi \mid X_i = x_i] + \lambda := 0$$

$$\ln Q(x_i) = -\lambda - 1 + \sum_{\phi\in\Phi} E_{\mathcal{X}\sim Q}[\ln\phi \mid X_i = x_i]$$

the distribution Q is a stationary point of mean-field optimization iff

$$Q(x_i) = \frac{1}{Z_i}\exp\Big\{\sum_{\phi\in\Phi} E_{\mathcal{X}\sim Q}[\ln\phi \mid X_i = x_i]\Big\}$$

$$\sum_{\phi\in\Phi} E_{\mathcal{X}\sim Q}[\ln\phi \mid X_i = x_i] = E_{\mathcal{X}\sim Q}\Big[\sum_{\phi\in\Phi}\ln\phi \mid X_i = x_i\Big] = E_{\mathcal{X}\sim Q}[\ln\tilde{P}_\Phi(x) \mid X_i = x_i]$$

$$= E_{X_{-i}\sim Q}[\ln\tilde{P}_\Phi(x_i, X_{-i})]$$ // Q is product of marginals, $X_{-i} = \{\mathcal{X} - X_i\}$

$$= E_{X_{-i}\sim Q}[\ln P_\Phi(x_i \mid X_{-i})] + E_{X_{-i}\sim Q}[\ln P_\Phi(X_{-i})\cdot Z]$$

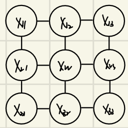
$$Q(x_i) = \frac{1}{Z_i}\exp\{E_{X_{-i}\sim Q}[\ln P_\Phi(x_i \mid X_{-i})]\}\underbrace{\exp\{E_{X_{-i}\sim Q}[\ln P(X_{-i})\,Z]\}}_{\text{does not depend on } x_i}$$

In the mean field approximation, $Q(X_i)$ is locally optimal only if

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

$Q(x_i)$ has to be consistent with the expectation of the potential in which it appears

$$Q(x_{ij}) = \frac{1}{Z_{ij}} \exp\Big\{ \sum_{x_{i-1,j}} Q(x_{i-1,j}) \cdot \ln \phi(x_{i-1,j}, x_{ij})$$
$$+ \sum_{x_{i,j-1}} Q(x_{i,j-1}) \cdot \ln \phi(x_{i,j-1}, x_{ij})$$
$$+ \sum_{x_{i+1,j}} Q(x_{i+1,j}) \ln \phi(x_{ij}, x_{i+1,j})$$
$$+ \sum_{x_{i,j+1}} Q(x_{i,j+1}) \ln \phi(x_{ij}, x_{i,j+1})$$
$$\Big\}$$



each term is a geometric mean of one potential involving X_{ij}

```
def mean_field_approximation(Φ, Q0):
    Q = Q0
    Unprocessed = X
    while Unprocessed ≠ ∅:
        choose X_i from Unprocessed
        Q_old(X_i) = Q(X_i)
        for x_i ∈ Val(X_i)
            Q(x_i) = exp{ Σ_{φ: X_i ∈ Scope[φ]} E_{{Uφ - X_i}~Q} [ ln φ(Uφ, x_i) ] }
        Normalize Q(X)
        if Q_old(X_i) ≠ Q(X_i)
            Unprocessed = Unprocessed ∪ ( ∪_{φ: X_i ∈ Scope[φ]} Scope[φ] )
        Unprocessed = Unprocessed - {X_i}
    return Q
```

Coordinate ascent