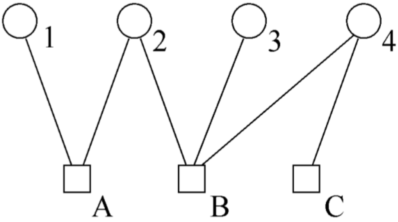
Outline

- Belief Propagation
- Bethe Method
- EP and Divergences

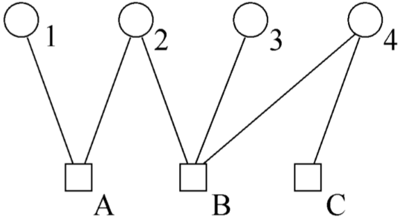
Factor Graph



$$p(x_1, x_2, x_3, x_4) = \frac{1}{Z} f_A(x_1, x_2) f_B(x_2, x_3, x_4) f_C(x_4) \qquad (1)$$

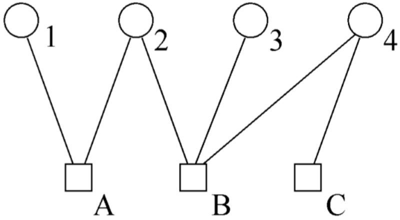
$$p(\mathbf{x}_S) = \sum_{\mathbf{x} \setminus \mathbf{x}_S} p(\mathbf{x}), \quad \forall S \subset \{x_1, x_2, x_3, x_4\}$$

Message Passing



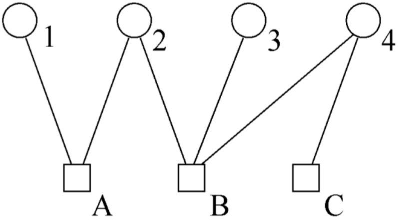
- message from factor to node:
  $M_{a\to i}(x_i) := \sum_{\mathbf{x}_a \setminus x_i} f_a(\mathbf{x}_a) \prod_{j\in N(a)\setminus i} M_{j\to a}(x_j)$
- message from node to factor:
  $M_{j\to a}(x_j) := \prod_{a'\in N(j)\setminus a} M_{a'\to j}(x_j)$

Message Passing



- belief (or pseudo prob.) of the node:
  $q_i(x_i) \propto \prod_{a \in N(i)} M_{a \to i}(x_i)$
- belief of the factor:
  $q_a(\mathbf{x}_a) \propto f_a(\mathbf{x}_a) \prod_{i \in N(a)} M_{i \to a}(x_i)$
- $q_i(x_i) = \sum_{\mathbf{x}_a \setminus x_i} q_a(\mathbf{x}_a)$

Why Loopy Belief Propagation



- $q_i(x_i) = p(x_i)$ if no loops in the graph
- The approximation by BP will be worse with more loops
- Loopy BP: region-based free energy approximations

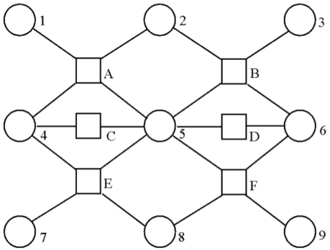
Free Energies

- Boltzmann's Law: $p(\mathbf{x}) = \frac{1}{Z} e^{-E(\mathbf{x})}$
  $E(\mathbf{x}) := -\sum_a \log f_a(\mathbf{x}_a)$
- Helmholtz free energy: $F_{Helmholtz} = -\log Z$
- Variational (or Gibbs) free energy:

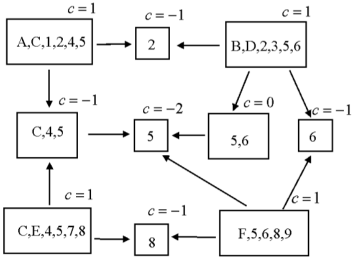
$$F(q) = \underbrace{\sum_{\mathbf{x}} q(\mathbf{x}) E(\mathbf{x})}_{U(q)} + \underbrace{\sum_{\mathbf{x}} q(\mathbf{x}) \log q(\mathbf{x})}_{-H(q)} \qquad (2)$$

- $F(q) = F_{Helmholtz} + KL(q||p)$

Region Graph



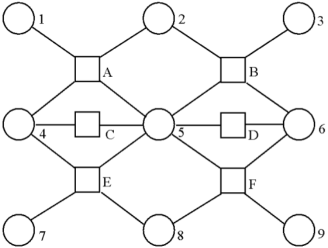
(a) factor graph



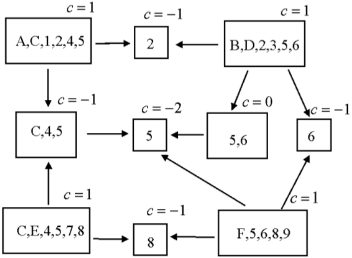
(b) region graph

- Recall a factor graph with vertices $I = \{i, a\}$
- A region graph is a labelled directed graph $\mathcal{G} = (V, E, L)$:
  - $v \in V$ is labelled by some subset $L(v) \subset I$
  - if $v_p \to v_c \in E$, then $L(v_c) \subset L(v_p)$
- A vertex $v \in V$ correspond to a region $R \subset I$

Region Graph



(c) factor graph

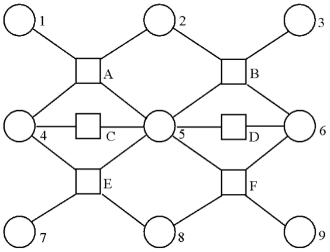


(d) region graph

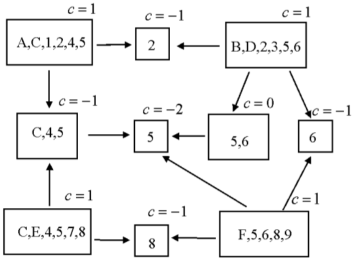
- Region energy: $E_R(\mathbf{x}_R) := -\sum_{a\in f(R)} \log f_a(\mathbf{x}_a)$
- Region free energy:

$$F_R(q_R) = \underbrace{\sum_{\mathbf{x}} q_R(\mathbf{x}_R) E_R(\mathbf{x}_R)}_{U_R(q_R)} + \underbrace{\sum_{\mathbf{x}_R} q_R(\mathbf{x}_R) \log q_R(\mathbf{x}_R)}_{-H_R(q_R)} \tag{3}$$

Region Graph



(e) factor graph



(f) region graph

- Define region count c_R (or c_v of corresponding vertex v): $\sum_{R \in \mathcal{R}} [a \in R] c_R = \sum_{R \in \mathcal{R}} [i \in R] c_R = 1$
- $F(q) = \sum_{R \in \mathcal{R}} c_R F_R(q_R)$
- $U(q) = \sum_{R \in \mathcal{R}} c_R U_R(q_R)$, $H(q) = \sum_{R \in \mathcal{R}} c_R H_R(q_R)$

Bethe Energy

- $\mathcal{R} = \mathcal{R}_L \bigcup \mathcal{R}_S$
  - $R \in \mathcal{R}_L$ only contains a factor node and its adjacent variable node
  - $R \in \mathcal{R}_S$ only contains one variable node
- $c_R = 1 - \sum_{S \in \mathcal{S}(R)} c_S$
  - $\mathcal{S}(R) = \{R' \in \mathcal{R} : L(R) \subset L(R')\}$
- $c_R = 1$, if $R \in \mathcal{R}_L$
- $c_R = 1 - d_i$, if $R \in \mathcal{R}_S$ contains variable i with degree d_i

Bethe Energy

- Bethe free energy: $F_{Bethe} = U_{Bethe} - H_{Bethe}$

$$\begin{aligned} U_{Bethe} &= -\sum_{a \in f(\mathcal{R})} \sum_{\mathbf{x}_a} q_a(\mathbf{x}_a) \log f_a(\mathbf{x}_a) \\ H_{Bethe} &= -\sum_{a \in f(\mathcal{R})} \sum_{\mathbf{x}_a} q_a(\mathbf{x}_a) \log q_a(\mathbf{x}_a) \\ &\quad + \sum_i (d_i - 1) \sum_{x_i} q_i(x_i) \log q_i(x_i) \end{aligned} \tag{4}$$

Bethe Approximation and Standard BP

Theorem

Let $\{M_{a\to i}(x_i), M_{i\to a}(x_i)\}$ be the BP messages and $\{q_a(\mathbf{x}_a), q_i(x_i)\}$ be the corresponding beliefs. Then the beliefs are fixed points of the BP algorithm iff. they are stationary points of the Bethe free energy F_{Bethe}.

- BP always has a fixed point
- Only one fixed point if there's no more than 1 cycle
- Exact approximation if no cycles in the factor graph:

$$p(\mathbf{x}) = \frac{\prod_i p_a(\mathbf{x}_a)}{\prod_i (p_i(x_i))^{d_i - 1}} = q(\mathbf{x}) \tag{5}$$

Bethe Method: Inference

- assume the single and pairwise potentials satisfy

$$p(\mathbf{x}) = \frac{1}{Z_p} \prod_{(ij)\in E} \psi_{ij}(x_i, x_j) \prod_i \psi_i(x_i)$$

  - define $\phi_{ij}(x_i, x_j) = \psi_{ij}(x_i, x_j)\psi_i(x_i)\psi_j(x_j)$
  - also define $\phi_i(x_i) = \psi_i(x_i)$
- Rewrite the Bethe energy

$$\begin{aligned} F_{Bethe} = &\sum_{(ij)\in E} \sum_{x_i, x_j} q_{ij}(x_i, x_j) \log \frac{q_{ij}(x_i, x_j)}{\phi_{ij}(x_i, x_j)} \\ &+ \sum_i (1 - n_i) \sum_{x_i} q_i(x_i) \log \frac{q_i(x_i)}{\phi_i(x_i)} \end{aligned} \tag{6}$$

Bethe Approximation

- Bethe approximation: minimize

$$F_{Bethe}(q) + \log Z_p \approx KL(q||p)$$

- Constraints of the approximation q:
  - observational constraint: $q(x_i) = \hat{o}_i(x_i)$
  - marginalization constraint: $\sum_{x_j} q_{ij}(x_i, x_j) = q_i(x_i)$
  - normalization constraint: $\sum_{x_i} q_i(x_i) = 1$
- The resulting Lagrangian

$$\mathcal{L} = F_{Bethe}(q) - \sum_i \sum_{j \in N(i)} \sum_{x_i} \lambda_{ji}(x_i) \left(\sum_{x_j} q_{ij}(x_i, x_j) - q_i(x_i) \right)$$

Bethe Approximation

Theorem

Subject to the constraints, the stationary points of F_{Bethe} is given by

$$q_{ij}(x_i, x_j) \propto \phi_{ij}(x_i, x_j) \exp(\lambda_{ji}(x_i) + \lambda_{ij}(x_j)) \tag{7}$$

$$q_i(x_i) \propto \phi_i(x_i) \exp(\frac{1}{d_i - 1} \sum_{j \in N(i)} \lambda_{ji}(x_i)) \tag{8}$$

where the Lagrange multipliers are fixed points of the following updates:

$$e^{\lambda_{ji}(x_i)} \leftarrow \prod_{k \in N(i) \setminus j} \sum_{x_k} \frac{\phi_{ik}(x_i, x_k)}{\phi_i(x_i)} e^{\lambda_{ik}(x_k)}, \quad \text{for hidden } i \tag{9}$$

$$e^{\lambda_{ji}(x_i)} \leftarrow \frac{\hat{o}_i(x_i)}{\sum_{x_i} \phi_{ij}(x_i, x_j) e^{\lambda_{ij}(x_j)}}, \quad \text{for observed } i \tag{10}$$

Bethe Approximation (Message Passing)

- Define messages $M_{i\to j}(x_j) = \sum_{x_i} \frac{\phi_{ij}(x_i,x_j)}{\phi_j(x_j)} e^{\lambda_{ji}(x_i)}$
- Rewrite (9)

$$e^{\lambda_{ji}(x_i)} \leftarrow \prod_{k\in N(i)\backslash j} M_{k\to i}(x_i), \quad \textit{for hidden } i \tag{11}$$

- Recover BP updates

$$M_{i\to j}(x_j) \leftarrow \sum_{x_i} \frac{\phi_{ij}(x_i,x_j)}{\phi_j(x_j)} \prod_{k\in N(i)\backslash j} M_{k\to i}(x_i) \tag{12}$$

- Can also rewrite (10)

$$M_{i\to j}(x_j) \leftarrow \sum_{x_i} \psi_{ij}(x_i,x_j) \frac{\hat{o}(x_i)}{M_{j\to i}(x_i)}, \quad \textit{for observed } i \tag{13}$$

Bethe Method: Learning

- Maximum entropy: given (empirical) marginals $\hat{p}$

$$q^* = \arg\max_q H(q) \quad s.t. \; q(\mathbf{x}_a) = \hat{p}(\mathbf{x}_a) \tag{14}$$

- Implementing constraints and the resulting Lagrangian

$$\mathcal{L} = H(q) - \sum_{a,\mathbf{x}_a} \lambda_a(\mathbf{x}_a)(\hat{p}(\mathbf{x}_a) - \sum_{\mathbf{x}_{\setminus a}} q(\mathbf{x})) - \gamma(1 - \sum_{\mathbf{x}} q(\mathbf{x})) \tag{15}$$

- Zeroing derivatives of $\mathcal{L}$ wrt. q and γ

$$q(\mathbf{x}) = \frac{1}{Z} e^{\sum_a \lambda_a(\mathbf{x}_a)} \tag{16}$$

Maximum Entropy

- Dual cost (convex)

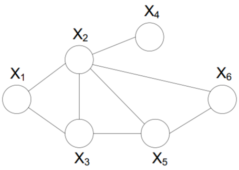
$$\mathcal{L}' = -\sum_{a,\mathbf{x}_a} \lambda_a(\mathbf{x}_a)\hat{p}(\mathbf{x}_a) + \log \sum_{\mathbf{x}} e^{\sum_a \lambda_a(\mathbf{x}_a)} \tag{17}$$

- Solved by coordinate-wise descent in λ_a

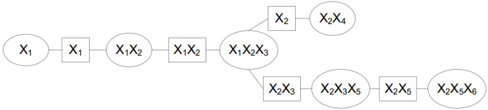
$$\lambda_a(\mathbf{x}_a) \leftarrow \lambda_a(\mathbf{x}_a) + \log \frac{\hat{p}(\mathbf{x}_a)}{q(\mathbf{x}_a)} \tag{18}$$

- Equivalent to $q(\mathbf{x}) \leftarrow q(\mathbf{x})\frac{\hat{p}(\mathbf{x}_a)}{q(\mathbf{x}_a)}$
- Equivalent to maximum likelihood

Junction Trees



(g) graphical model



(h) junction tree

- Formed by maximal cliques C with separators S

$$q(\mathbf{x}) = \frac{\prod_{c \in C} q_c(\mathbf{x}_c)}{\prod_{s \in S} q_s(\mathbf{x}_s)} \tag{19}$$

- For any cluster a, there exist $c \in C$ s.t. $a \subset c$
- $q_{c_1}(\mathbf{x}_s) = q_{c_2}(\mathbf{x}_s)$ if c_1, c_2 are neighbouring cliques separated by s

Junction Trees

- Learning by maximum entropy

$$\arg\max_{\{q_c, q_s\}} \sum_c H(q_c) - \sum_s H(q_s) \tag{20}$$

subject to $q_c(\mathbf{x}_a) = \hat{p}_a(\mathbf{x}_a), \; q_c(\mathbf{x}_s) = q_s(\mathbf{x}_s), \; \forall a, s \subset c$

- The resulting Lagrangian (def. $a \subset c_a$)

$$\begin{aligned}\mathcal{L} = & \sum_c H(q_c) - \sum_s H(q_s) - \sum_{v \in S \cup C} \gamma_v (\sum_{\mathbf{x}_v} q_v(\mathbf{x}_v) - 1) \\ & - \sum_{c,s,\mathbf{x}_s} \lambda_{cs}(\mathbf{x}_s)(q_s(\mathbf{x}_s) - \sum_{\mathbf{x}_{c \setminus s}} q_c(\mathbf{x}_c)) \\ & - \sum_{a,\mathbf{x}_a} \lambda_a(\mathbf{x}_a)(\hat{p}_a(\mathbf{x}_a) - \sum_{\mathbf{x}_{c_a \setminus a}} q_{c_a}(\mathbf{x}_{c_a}))\end{aligned} \tag{21}$$

Junction Trees

Theorem

Define $A_c := \{a | c_a = c\}$. *Then solving the Lagrangian returns marginal distributions*

$$q_c(\mathbf{x}_c) \propto e^{\sum_s \lambda_{cs}(\mathbf{x}_s) + \sum_{a \in A_c} \lambda_a(\mathbf{x}_a)} \tag{22}$$

$$q_s(\mathbf{x}_s) \propto e^{\sum_c \lambda_{cs}(\mathbf{x}_s)} \tag{23}$$

while λ_a *and* λ_{cs} *are the fixed points of the following updates*

$$\lambda_a(\mathbf{x}_a) \leftarrow \lambda_a(\mathbf{x}_a) + \log \frac{\hat{p}_a(\mathbf{x}_a)}{q_{c_a}(\mathbf{x}_{c_a})} \tag{24}$$

$$e^{\lambda_{c's}} \leftarrow\propto \sum_{\mathbf{x}_{c \setminus s}} e^{\sum_{s' \neq s} \lambda_{cs'}(\mathbf{x}_{s'}) + \sum_a \lambda_{ca}(\mathbf{x}_a)} \tag{25}$$

where c', c *are separated by* s, *and* s' *are other separators neighbouring* c.

Junction Trees (Message Passing)

- Define messages and potentials (factors)

$$M_{c \to s}(x_s) := e^{\lambda_{cs}(\mathbf{x}_s)}, \quad f_c(\mathbf{x}_c) := e^{\sum_a \lambda_{ca}(\mathbf{x}_a)}$$

- (25) is equivalent to

$$M_{c' \to s}(x_s) \leftarrow\propto \sum_{\mathbf{x}_{c \setminus s}} f_c(\mathbf{x}_c) \prod_{s' \neq s} M_{c \to s'}(x_{s'}) \qquad (26)$$

- Rewrite the marginals (22) and (23)

$$q_c(\mathbf{x}_c) \propto f_c(\mathbf{x}_c) \prod_s M_{c \to s}(x_s), \quad q_s(\mathbf{x}_s) \propto \prod_c M_{c \to s}(x_s)$$

- ... or by Hugin propagation

$$q_{c'}(\mathbf{x}_{c'}) \leftarrow q_{c'}(\mathbf{x}_{c'}) \frac{q_c(\mathbf{x}_s)}{q_s(\mathbf{x}_s)}, \quad q_s(\mathbf{x}_s) \leftarrow q_c(\mathbf{x}_s)$$

EP energy

- EP approximation

$$p(\mathbf{x}|D) = p(\mathbf{x}) \prod_i^n t_i(\mathbf{x}) \approx p(\mathbf{x}) \prod_i \tilde{t}_i(\mathbf{x}) := q(\mathbf{x}) \quad (27)$$

- $\tilde{t}_i(\mathbf{x}) = e^{\sum_j f_j(\mathbf{x})\tau_j}$
- Minimizing (local) KL-divergence $KL(\hat{p}_i||q)$ where $\hat{p}_i := q^{\setminus i} t_i$
- ... by matching the expetations $E_{\hat{p}_i}[f_j]$ and $E_q[f_j]$
- May want q and $\hat{p}_i$ to be normalised

EP energy

- The EP primal energy function (satisfying moment maching and normalization constraints)

$$\min_{\hat{p}_i} \max_{q} \sum_{i}^{n} KL(\hat{p}_i || t_i p) - (n-1) KL(q||p) \tag{28}$$

- (Dual) energy function

$$\begin{aligned} \min_{\nu} \max_{\lambda} \; & (n-1) \log \int_{\mathbf{x}} p(\mathbf{x}) e^{\sum_j f_j(\mathbf{x}) \nu_j} d\mathbf{x} \\ & - \sum_{i}^{n} \log \int_{\mathbf{x}} t_i(\mathbf{x}) p(\mathbf{x}) e^{\sum_j f_j(\mathbf{x}) \lambda_{ij}} d\mathbf{x} \end{aligned} \tag{29}$$

s.t. $(n-1)\nu_j = \sum_i \lambda_{ij}$

Equavalence between BP and Bethe Energies

- BP is a special case of EP that f_j are delta functions
- Recall the Bethe energy

$$F_{Bethe} = \underbrace{\sum_{(ij)\in E} \sum_{x_i, x_j} q_{ij}(x_i, x_j) \log \frac{q_{ij}(x_i, x_j)}{\phi_{ij}(x_i, x_j)}}_{①}$$

$$\underbrace{- \sum_i (n_i - 1) \sum_{x_i} q_i(x_i) \log \frac{q_i(x_i)}{\phi_i(x_i)}}_{②}$$

- ... minimizing F_{Bethe} is by updating

$$q_{ij}(x_i, x_j) \propto \phi_{ij}(x_i, x_j) \exp(\lambda_{ji}(x_i) + \lambda_{ij}(x_j))$$

Equavalence between BP and Bethe Energies

- Another representation of the KL-divergence

$$KL(P||Q) = \max_{\nu} E_P[\nu(x)] - \log E_Q[e^{\nu(x)}] \tag{30}$$

- Apply to the Bethe energy

$$\begin{aligned} ① = \max_{\lambda} \sum_{x_i} q_i(x_i)\lambda_{ji}(x_i) + \sum_{x_j} q_j(x_j)\lambda_{ij}(x_j) \\ - \log \sum_{x_i,x_j} \phi_{ij}(x_i, x_j) e^{\lambda_{ji}(x_i)+\lambda_{ij}(x_j)} \end{aligned} \tag{31}$$

$$② = \min_{\nu} - \sum_{i} (n_i - 1) \sum_{x_i} q_i(x_i)\nu(x_i) + \log \sum_{x_i} \phi_i(x_i) e^{\nu(x_i)} \tag{32}$$

Equavalence between BP and Bethe Energies

- Substitute (31), (32) into $\min_q F_{Bethe}$ and zeroing the gradient wrt. ν and λ:

$$q_i(x_i) = \frac{\phi_i(x_i)e^{\nu(x_i)}}{Z_1} = \frac{\sum_{x_j} \phi_{ij}(x_i, x_j) e^{\lambda_{ji}(x_i)+\lambda_{ij}(x_j)}}{Z_2} \tag{33}$$

- Add constraint $(n_i - 1)\nu(x_i) = \sum_j \lambda_{ji}(x_i)$ to delete $q_i(x_i)$, then have the transformed objective

$$\begin{aligned} \min_\nu \max_\lambda \sum_i (n_i - 1) \log \sum_i \phi_i(x_i) e^{\nu(x_i)} \\ - \sum_{(ij)\in E} \log \sum_{x_i, x_j} \phi_{ij}(x_i, x_j) e^{\lambda_{ji}(x_i)+\lambda_{ij}(x_j)} \end{aligned} \tag{34}$$

Bethe Approximation, BP and EP

- Recall the coincidence of BP fixed points and Bethe energy stationary points
- EP extends BP
- EP fixed points = stationary points of some free energy function

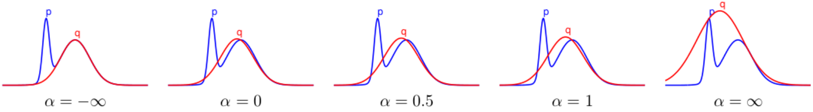
Power EP and α-Divergences

- Power EP: minimizing (local) KL-divergence $KL(q(\frac{t_i}{\tilde{t}_i})^\alpha || q)$
- Equivalent to minimize the α-divergence

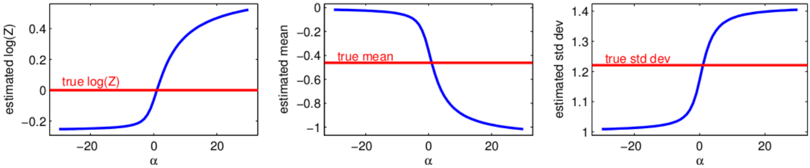
$$D_\alpha(\hat{p}_i || q) := \frac{\int_{\mathbf{x}} \alpha \hat{p}_i(\mathbf{x}) + (1-\alpha) q(\mathbf{x}) - \hat{p}_i(\mathbf{x})^\alpha q(\mathbf{x})^{(1-\alpha)} d\mathbf{x}}{\alpha(1-\alpha)} \tag{35}$$

- $\lim_{\alpha \to 0} D_\alpha(p||q) = KL(q||p)$
- $\lim_{\alpha \to 1} D_\alpha(p||q) = KL(p||q)$

Power EP and α-Divergences



(i) The Gaussian q which minimizes α-divergence to p (a mixture of two Gaussians)



(j) The mass, mean, and standard deviation of the Gaussian q which minimizes α-divergence to p