ECE 8803

Approximate Inference in Graphical Models

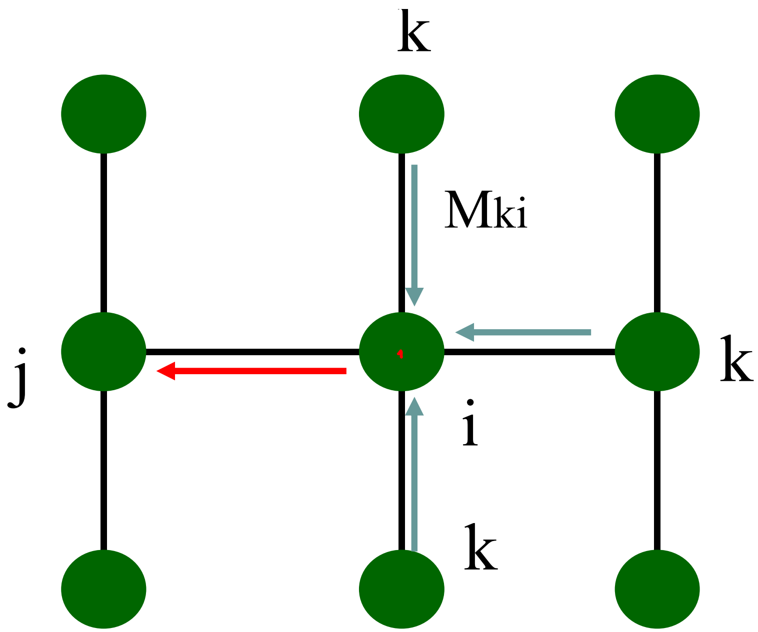
Module 8: Part B

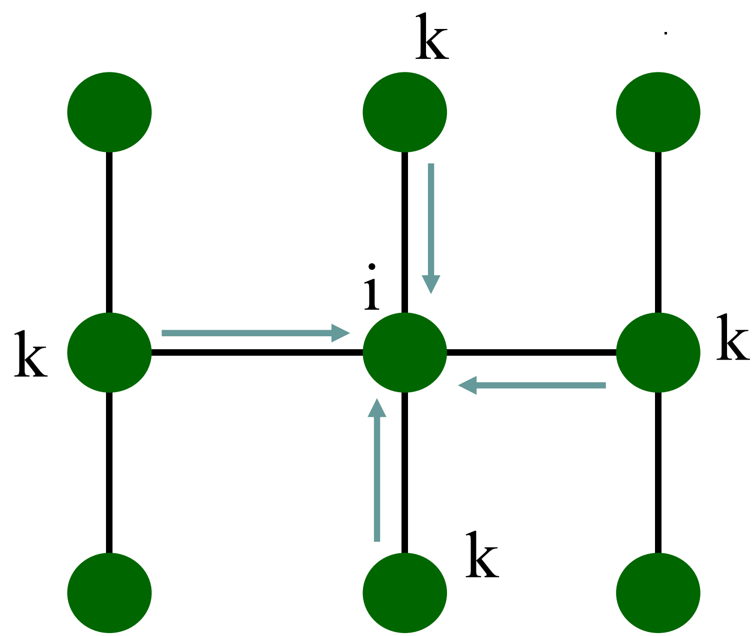
Variational Inference via Loopy Belief Propagation

Faramarz Fekri

Center for Signal and Information Processing

Recall: Belief Propagation





BP Message-update Rules

$$M_{i\to j}(x_j) \propto \sum_{x_i} \psi_{ij}(x_i, x_j)\psi_i(x_i)\prod_k M_{k\to i}(x_i)$$

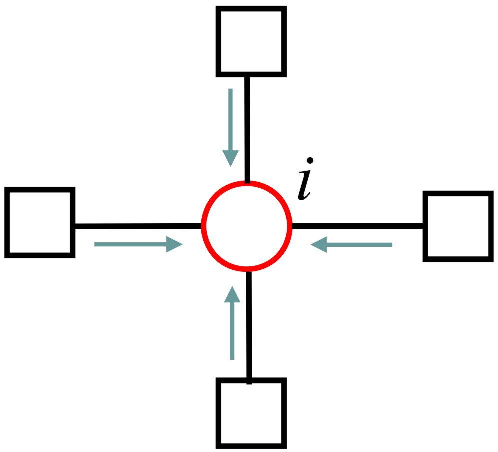
$$b_i(x_i) \propto \psi_i(x_i)\prod_k M_k(x_k)$$

external evidence

Compatibilities (interactions)

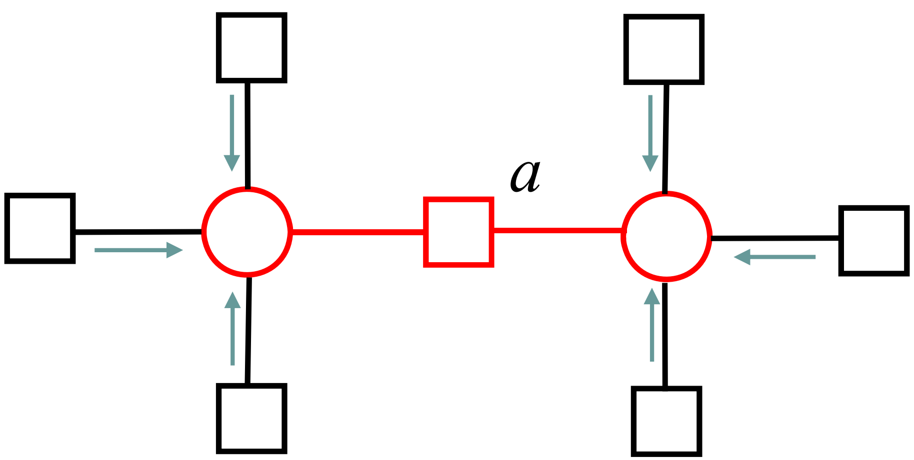
BP on trees always converges to exact marginals (e.g., Junction tree algorithm)

Beliefs and Messages in Factor Graph



$$b_i(x_i) \propto f_i(x_i) \prod_{a \in N(i)} m_{a \to i}(x_i)$$

"beliefs" "messages"

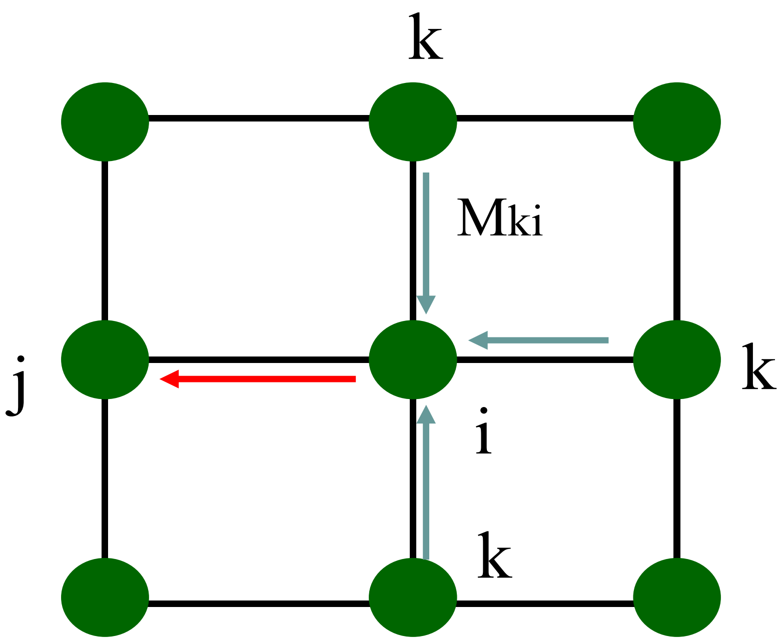


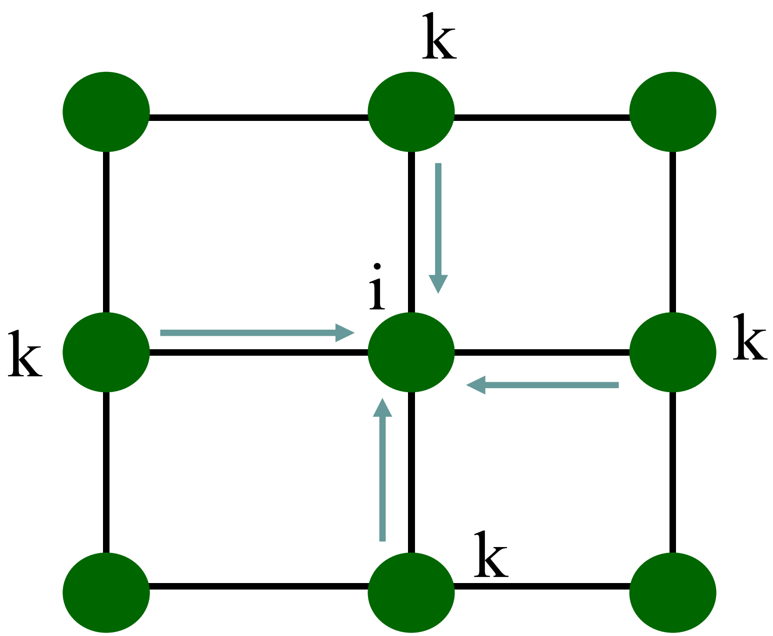
$$m_{i \to a}(x_i) = \prod_{c \in N(i) \backslash a} m_{c \to i}(x_i)$$

$$b_a(X_a) \propto f_a(X_a) \prod_{i \in N(a)} m_{i \to a}(x_i)$$

$$m_{a \to i}(x_i) = \sum_{X_a \backslash x_i} f_a(X_a) \prod_{j \in N(a) \backslash i} m_{j \to a}(x_j)$$

Belief Propagation on Loopy Graphs





- BP Message-update Rules

$$M_{i\to j}(x_j) \propto \sum_{x_i} \psi_{ij}(x_i, x_j)\psi_i(x_i)\prod_k M_{k\to i}(x_i)$$

$$b_i(x_i) \propto \psi_i(x_i)\prod_k M_k(x_k)$$

external evidence

Compatibilities (interactions)

- May not converge or converge to a wrong solution

What if we don’t have pairwise Markov nets? Transform to a pairwise MN

BP on Loopy Factor Graph

- Start with random initialization of messages and beliefs
  - While not converged do

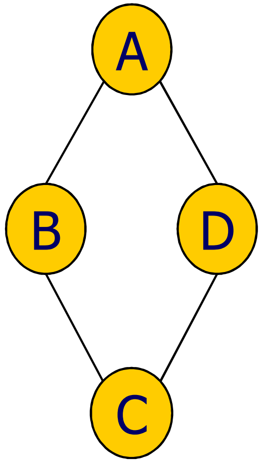
$$b_i(x_i) \propto \prod_{a \in N(i)} m_{a \to i}(x_i) \qquad b_a(X_a) \propto f_a(X_a) \prod_{i \in N(a)} m_{i \to a}(x_i)$$

$$m_{i \to a}^{new}(x_i) = \prod_{c \in N(i) \backslash a} m_{c \to i}(x_i) \qquad m_{a \to i}^{new}(x_i) = \sum_{X_a \backslash x_i} f_a(X_a) \prod_{j \in N(a) \backslash i} m_{j \to a}(x_j)$$

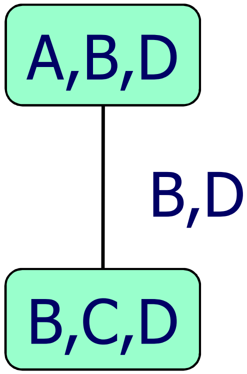
  - At convergence, stationarity properties are guaranteed
  - However, not guaranteed to converge!
- Empirically, a good approximation is still achievable
  - Stop after fixed # of iterations
  - Stop when no significant change in beliefs
  - If solution is not oscillatory but converges, it usually is a good approximation

A fixed point iteration procedure that tries to minimize F_{bethe}

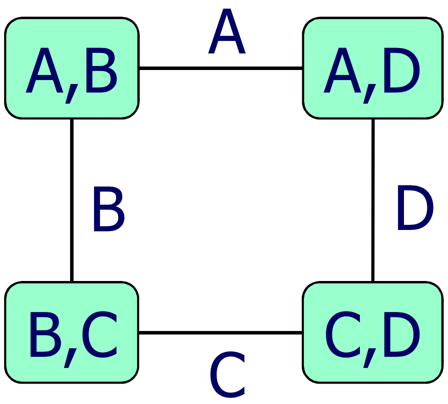
BP in a Cluster Graph with Loops



Simple network



Clique tree



Cluster graph

$$m_{D_j D_i}(S_{ji}) \propto \sum_{D_j \setminus S_{ji}} \Phi(D_j) \prod_{D_t \in \mathrm{N}(D_j) \setminus D_i} m_{D_t D_j}(S_{tj})$$

D_i: cluster i in the cluster graph

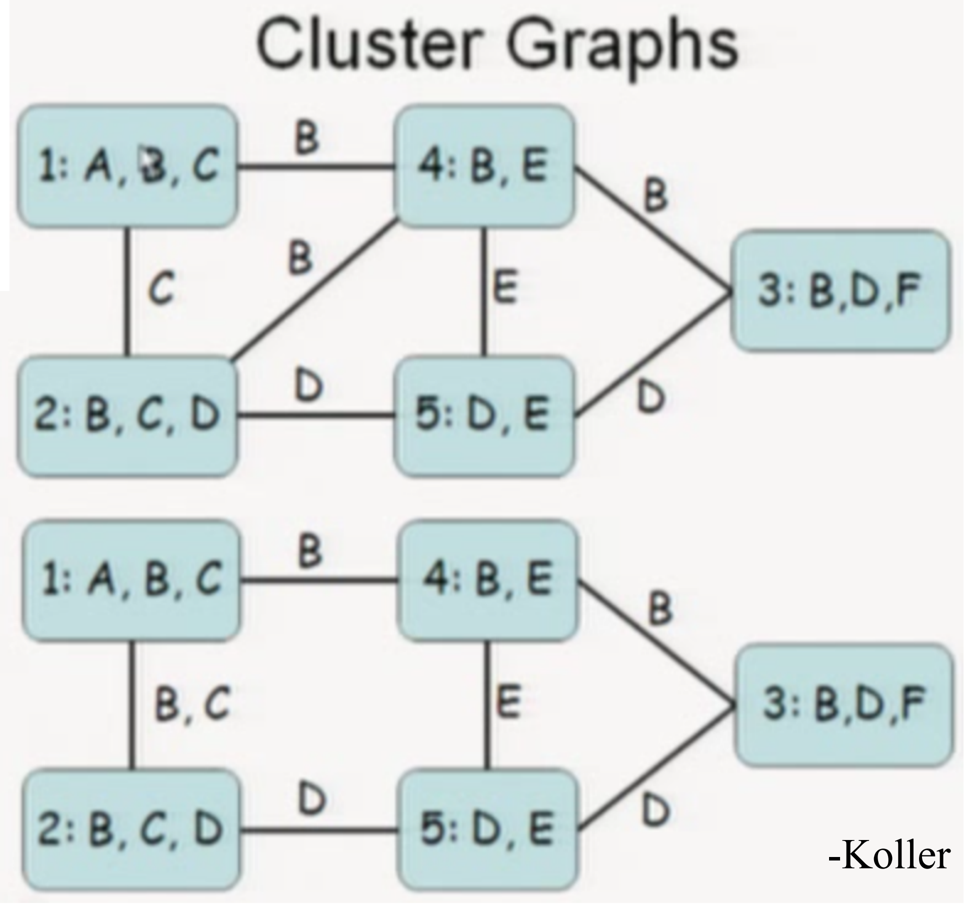
In Loopy BP, different cluster graphs can vary in both computational complexity and approximation quality (accuracy).

Desirable Cluster Graphs for Loopy BP

A generalized cluster graph K for factors F is an undirected graph

- Nodes are associated with a subset of variables $\mathbf{C_i} \subseteq \mathbf{U}$
- The graph is family preserving: each factor $\phi \in F$ is associated with one node $\mathbf{C_i}$ such that Scope$[\phi] \subseteq \mathbf{C_i}$
- Each edge $\mathbf{C_i}$–$\mathbf{C_j}$ is associated with a subset $\mathbf{S_{i,j}} \subseteq \mathbf{C_i} \cap \mathbf{C_j}$

- A generalized cluster graph obeys the running intersection property if for each $X \in C_i$ and $X \in C_j$, there is exactly one path between C_i and C_j for which $X \in S$ for each subset S along the path.
- All edges associated with X form a tree that spans all the clusters that contain X.
- Note: some of these clusters may be connected with more than one path.

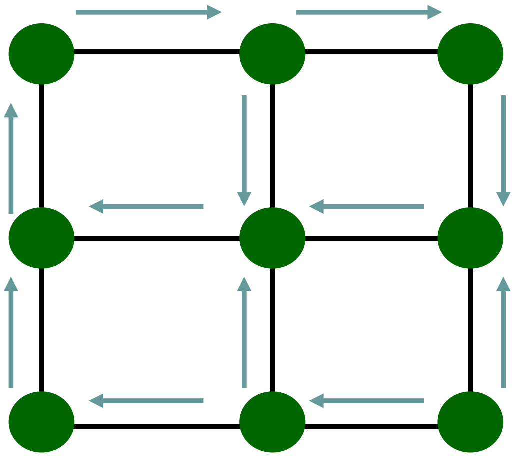


Lower graph is more desirable in case B and C are highly coupled since the upper graph will have implicit running intersection property in a loop.

Loopy Belief Propagation on Factor Graph

- What is going on when we ran Loopy BP?
- Let focus on Loopy BP on factor graphs (similar conclusion exists for BP over loopy cluster graphs)

$$KL(Q \| P) = \underbrace{\boxed{-H_Q(X) - \sum_{f_a \in F} E_Q \log f_a(X_a)}}_{F(P,Q)} + \log Z$$



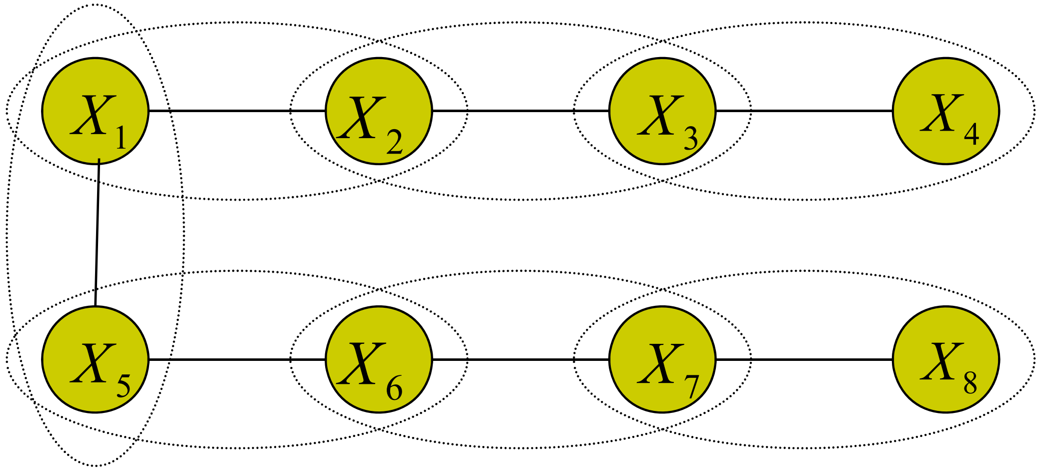
Note that the "(Gibbs) Free energy" here in loopy BP is minus of the free energy formulation we had in Mean Field, ie., we are minimizing F rather than maximizing it

- Energy functional: $F(P,Q) = -H_Q(X) - \sum_{f_a \in F} E_Q \log f_a(X_a)$

Approach: Approximate $F(P,Q)$ with easy to compute $F(P,Q)$

Tree Energy Functionals

- Consider a tree-structured distribution



- The probability can be written as: $b(\mathbf{x}) = \prod_a b_a(\mathbf{x}_a) \prod_i b_i(x_i)^{1-d_i}$
- $$H_{tree} = -\sum_a \sum_{\mathbf{x}_a} b_a(\mathbf{x}_a) \ln b_a(\mathbf{x}_a) + \sum_i (d_i - 1) \sum_{\mathbf{x}_i} b_i(\mathbf{x}_i) \ln b_i(\mathbf{x}_i)$$
- $$F_{Tree} = \sum_a \sum_{\mathbf{x}_a} b_a(\mathbf{x}_a) \ln \frac{b_a(\mathbf{x}_a)}{f_a(\mathbf{x}_a)} + \sum_i (1 - d_i) \sum_{\mathbf{x}_i} b_i(\mathbf{x}_i) \ln b_i(\mathbf{x}_i)$$
  $$= F_{12} + F_{23} + .. + F_{67} + F_{78} - F_1 - F_5 - F_2 - F_6 - F_3 - F_7$$
  - involves summation over edges and vertices and is therefore easy to compute

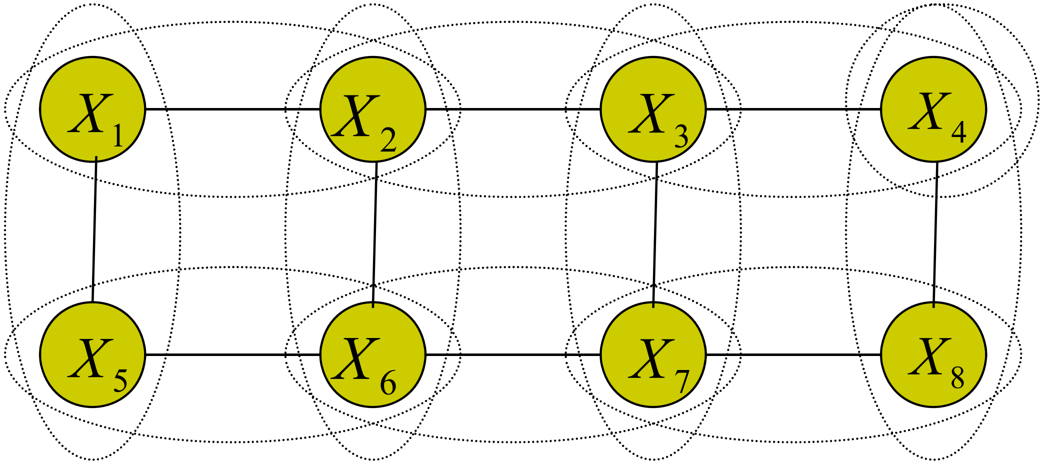
Bethe Approximation to Gibbs Free Energy

- For a general graph, choose $\hat{F}(P,Q) = F_{Betha}$

$$H_{Bethe} = -\sum_a \sum_{\mathbf{x}_a} b_a(\mathbf{x}_a) \ln b_a(\mathbf{x}_a) + \sum_i (d_i - 1) \sum_{\mathbf{x}_i} b_i(\mathbf{x}_i) \ln b_i(\mathbf{x}_i)$$

$$F_{Bethe} = \sum_a \sum_{\mathbf{x}_a} b_a(\mathbf{x}_a) \ln \frac{b_a(\mathbf{x}_a)}{f_a(\mathbf{x}_a)} + \sum_i (1 - d_i) \sum_{\mathbf{x}_i} b_i(\mathbf{x}_i) \ln b_i(\mathbf{x}_i) = -\langle f_a(\mathbf{x}_a) \rangle - H_{betha}$$

  - Called "Bethe approximation" after the physicist Hans Bethe



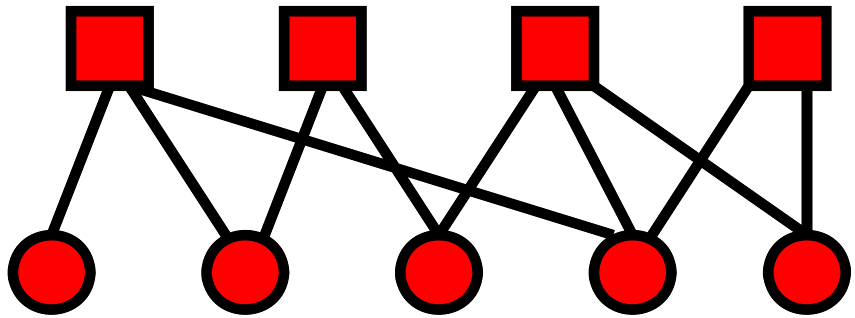
$$F_{bethe} = F_{12} + F_{23} + .. + F_{67} + F_{78} - F_1 - F_5 - 2F_2 - 2F_6 .. - F_8$$

  - Equal to the exact Gibbs free energy when the factor graph is a tree
  - In general, H_{Bethe} is **not** the same as the H of a tree

Bethe Approximation

- Pros:
  - Easy to compute, since entropy term involves sum over pairwise and single variables
- Cons:
  - $\hat{F}(P,Q) = F_{bethe}$ **may or may not** be well connected to $F(P,Q)$
  - It could, in general, be greater, equal or less than $F(P,Q)$
- Optimize each $b(\mathbf{x}_a)$'s.
  - For discrete belief, constrained opt. with *Lagrangian* multiplier
  - For continuous belief, not yet a general formula
  - Not always converge

Bethe Free Energy for Factor Graph



$$F_{Betha} = \sum_a \sum_{\mathbf{x}_a} b_a(\mathbf{x}_a) \ln \frac{b_a(\mathbf{x}_a)}{f_a(\mathbf{x}_a)} + \sum_i (1 - d_i) \sum_{\mathbf{x}_i} b_i(\mathbf{x}_i) \ln b_i(\mathbf{x}_i)$$

$$H_{Bethe} = -\sum_a \sum_{\mathbf{x}_a} b_a(\mathbf{x}_a) \ln b_a(\mathbf{x}_a) + \sum_i (d_i - 1) \sum_{\mathbf{x}_i} b_i(\mathbf{x}_i) \ln b_i(\mathbf{x}_i)$$

$$F_{Bethe} = -\langle f_a(\mathbf{x}_a) \rangle - H_{betha}$$

Constrained Minimization of the Bethe Free Energy

$$L = F_{Bethe} + \sum_i \gamma_i \{ \sum_{x_i} b_i(x_i) - 1 \}$$

$$+ \sum_a \sum_{i \in N(a)} \sum_{x_i} \lambda_{ai}(x_i) \left\{ \sum_{X_a \setminus x_i} b_a(X_a) - b_i(x_i) \right\}$$

$$\frac{\partial L}{\partial b_i(x_i)} = 0 \quad \Longrightarrow \quad b_i(x_i) \propto \exp\left(\frac{1}{d_i - 1} \sum_{a \in N(i)} \lambda_{ai}(x_i) \right)$$

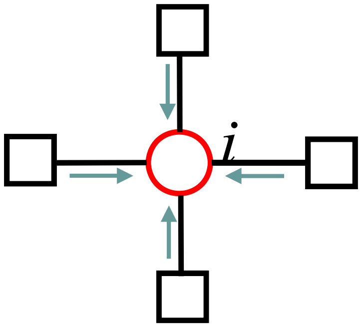
$$\frac{\partial L}{\partial b_a(X_a)} = 0 \quad \Longrightarrow \quad b_a(X_a) \propto \exp\left(-E_a(X_a) + \sum_{i \in N(a)} \lambda_{ai}(x_i) \right)$$

Minimization of Bethe Energy = Loopy BP on FG

- We had:

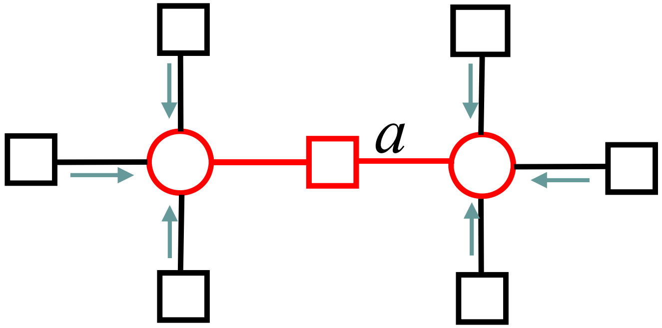
$$b_i(x_i) \propto \exp\left(\frac{1}{d_i - 1}\sum_{a\in N(i)}\lambda_{ai}(x_i)\right) \qquad b_a(X_a) \propto \exp\left(-\log f_a(X_a) + \sum_{i\in N(a)}\lambda_{ai}(x_i)\right)$$

- Identify $\lambda_{ai}(x_i) = \log(m_{i\to a}(x_i)) = \log \prod_{b\in N(i)\neq a} m_{b\to i}(x_i)$
- to obtain BP equations:



$$b_i(x_i) \propto f_i(x_i)\prod_{a\in N(i)} m_{a\to i}(x_i)$$

"beliefs" "messages"



$$b_a(X_a) \propto f_a(X_a)\prod_{i\in N(a)}\ \prod_{c\in N(i)\backslash a} m_{c\to i}(x_i)$$

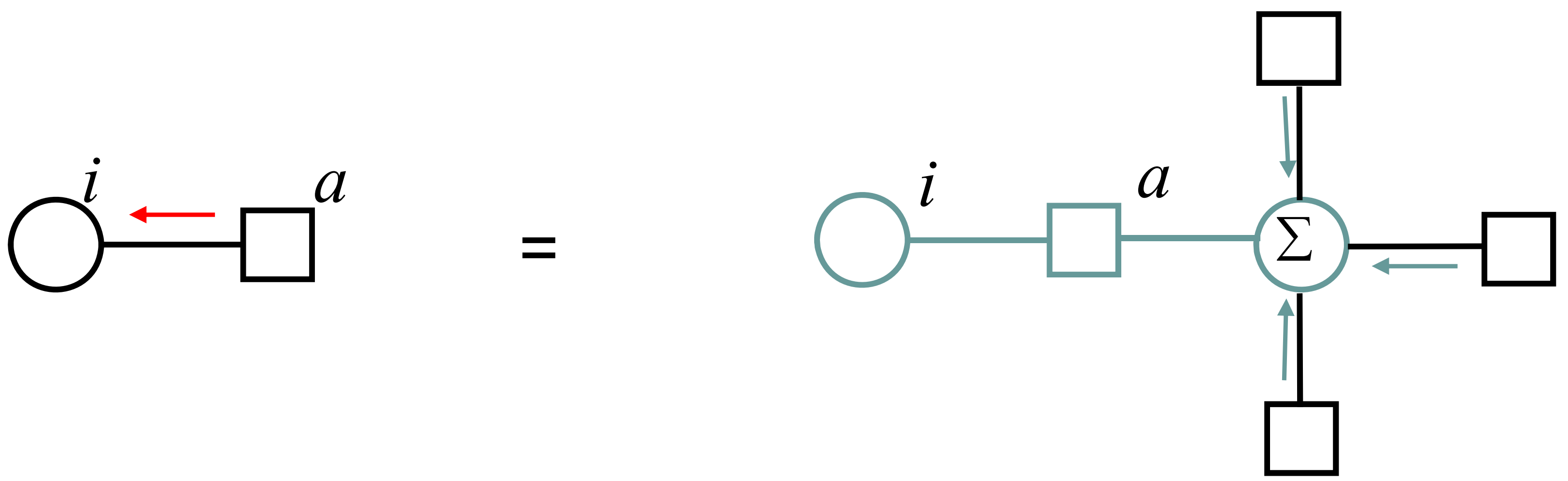
The "belief" is the BP approximation of the marginal probability.

BP Message-update Rules

Using $b_{a\to i}(x_i) = \sum_{X_a \backslash x_i} b_a(X_a)$, we get

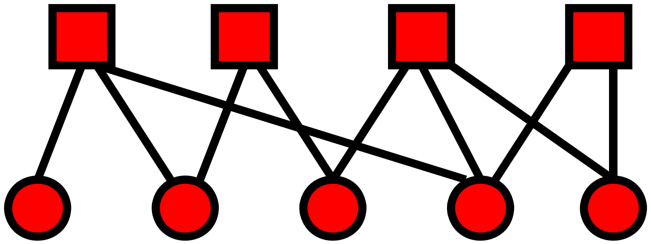
$$\boxed{m_{a\to i}(x_i) = \sum_{X_a \backslash x_i} f_a(X_a) \prod_{j\in N(a)\backslash i} \prod_{b\in N(j)\backslash a} m_{b\to j}(x_j)}$$

(A sum product algorithm)



Conclusion

- $$P(X)=1/Z\prod_{f_a\in F} f_a(X_a)$$

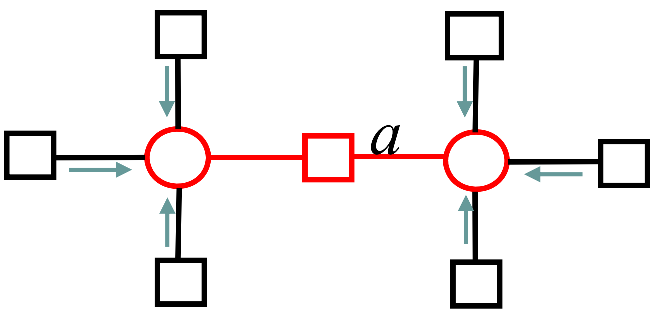


$$F(P,Q)=-H_Q(X)-\sum_{f_a\in F}E_Q\log f_a(X_a)$$

$$\hat{F}(P,Q)=\sum_a\sum_{\mathbf{x}_a}b_a(\mathbf{x}_a)\log\frac{f_a(\mathbf{x}_a)}{b_a(\mathbf{x}_a)}+\sum_i(1-d_i)\sum_{\mathbf{x}_i}b_i(\mathbf{x}_i)\log b_i(\mathbf{x}_i)$$

$$b_a(X_a)\propto\exp\left(-\log f_a(X_a)+\sum_{i\in N(a)}\lambda_{ai}(x_i)\right)$$

$$b_i(x_i)\propto\exp\left(\frac{1}{d_i-1}\sum_{a\in N(i)}\lambda_{ai}(x_i)\right)$$



Acknowledgement

- The materials of the lecture was mostly from Xing.