Free Energy Approximation

Solmaz Torabi

Dept. of Electrical and Computer Engineering
Drexel University
st669@drexel.edu

Advisor: Dr. John M. Walsh

Refrences

- M. Opper and D. Saad, "Advanced mean field methods: Theory and practice," *MIT press*, 2001.
- J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms." *Information Theory, IEEE Transactions*, vol. 51, 2005.
- M. Welling and Y. W. Teh, "Approximate inference in boltzmann machines," *Artificial Intelligence*, vol. 143, pp. 19–50, 2003.
- A. Montanari, "Lecture notes, inference in graphical models," 2011.

Outline

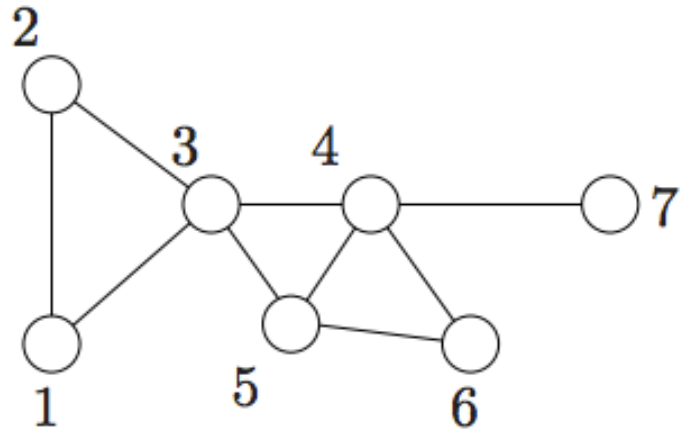
- Basics of graphical model
- Basics of message passing algorithm
- Variational free energy
  - Mean field approximation
  - TAP (Thouless, Anderson and Palmer)
  - Region Based approximation
    - Bethe free energy
    - Kikuchi approximation

Undirected graphical model, Markov random field

Undirected graphical model with random vector $X = (X_1, ..., X_n)$

- Given an undirected graph $G = (V, E)$, each node s has an associated random variable X_s
- A clique $C \subseteq V$ is a fully connected subset of V.
- The distribution p factorizes according to G if it can be expressed as a product over cliques.

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C)$$

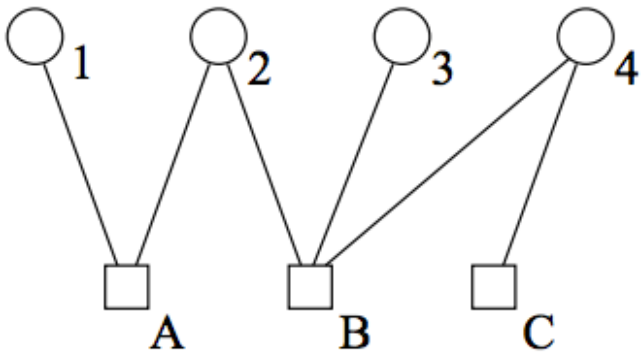


$$p(x) = \frac{1}{Z} \psi_1(x_1, x_2, x_3) \psi_2(x_3, x_4, x_5) \psi_3(x_4, x_5, x_6) \psi_4(x_4, x_7)$$

graphical model, Factor Graph

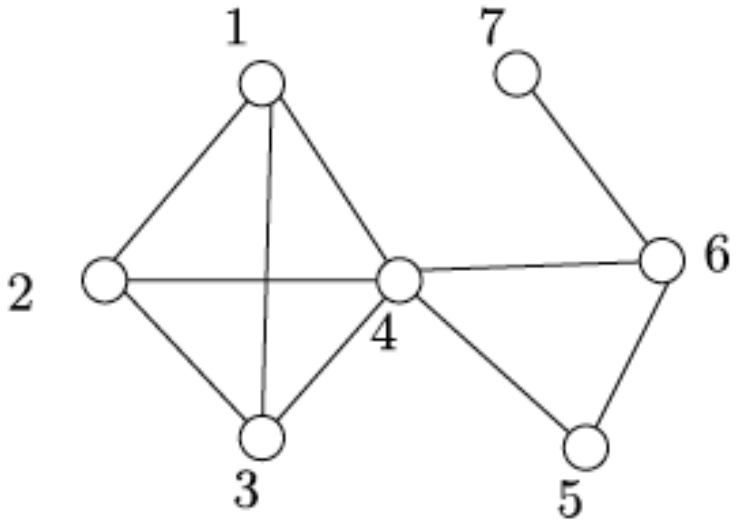
- Factor graph is bipartite graph $G = (V, F, E)$, where V is the original set of vertices, and $(s, a) \in E$ if x_s participates in the factor indexed by $a \in F$
- We assume that the functions $f_a(x_a)$ are non-negative and finite.

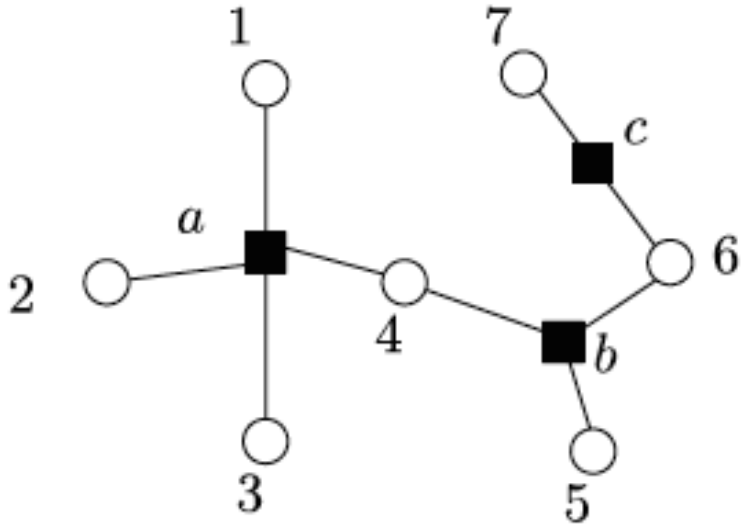
$$P(\mathbf{X}) = \frac{1}{Z} \prod_a f_a(x_\mathbf{a})$$



$$P(\mathbf{x}) = \frac{1}{Z} f_A(x_1, x_2) f_B(x_2, x_3, x_4) f_C(x_4)$$

graphical model- Undirected graph, Factor Graph





- Maximal cliques:
  $\mathcal{C} = \{1, 2, 3, 4\}, \{4, 5, 6\}, \{6, 7\}$
- Vertex set $V = \{1, ..., 7\}$
  factor set $F = \{a, b, c\}$

$$P(\mathbf{x}) = \frac{1}{Z} f_a(x_1, x_2, x_3, x_4) f_b(x_4, x_5, x_6) f_c(x_6, x_7)$$

Pairwise graphical model

- Subclass of Markov networks commonly encountered
  - Ising model, Boltzmann machines
  - Computer vision

$$P(x_1, x_2, ...x_N) = \frac{1}{Z} \prod_{(ij)} \psi_{ij}(x_i, x_j) \prod_i \psi_i(x_i)$$

where $\psi_{ij}(x_i, x_j)$ is compatibility function and $\psi_i(x_i)$ is the evidence of node i
$\psi_i : \mathcal{X} \to R_+$ for each $i \in \mathcal{V}$
$\psi_{ij} : \mathcal{X} \times \mathcal{X} \to R_+$ for each $(i, j) \in \mathcal{E}$

Boltzmann distribution

- Physicists specialize on the class of distribution P known as Boltzman distribution (Gibbs distribution)

$$P(\mathbf{X}) = \frac{e^{-H[\mathbf{X}]}}{Z}$$

  - $H(\mathbf{X})$ is the energy of each state
  - $Z = \sum_{\mathbf{X}} e^{-H[\mathbf{X}]}$ is the normalizing partition function
  - Pair-wise Markov random Field

$$P(\mathbf{X}) = \frac{1}{Z} \prod_{(ij)} \psi_{ij}(x_i, x_j) \prod_i \psi_i(x_i) = \frac{e^{-H[\mathbf{X}]}}{Z}$$

energy is

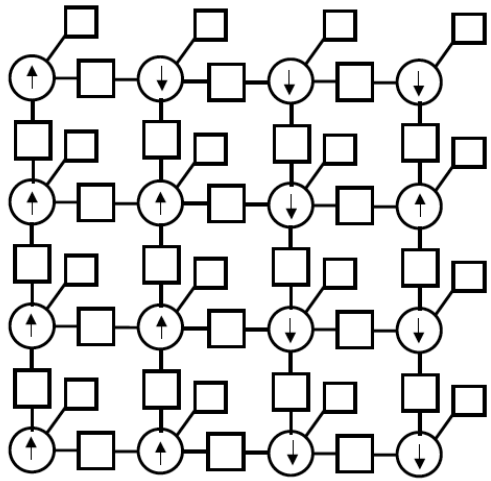
$$H[\mathbf{X}] = -\sum_{ij} \ln \psi_{ij}(x_i, x_j) - \sum_i \ln \psi_i(x_i)$$

Ising model

- An example of pairwise model with $\psi_{ij}(x_i, x_j) = \exp\{J_{ij}x_i x_j\}$, $\psi_i(x_i) = \exp\{\theta_i x_i\}$
- is a mathematical model of ferromagnetism in statistical mechanics.
- x_i represents magnetic dipole moments of atomic spins, $x_i \in \{+1, -1\}$, any two adjacent sites i, j has an interaction J_{ij}
- each site i has an external magnetic field θ_i

- The energy for each configuration is

$$H(\mathbf{X}) = -\sum_{i,j} J_{ij}x_i x_j - \sum_i \theta_i x_i$$



- The configuration probability is

$$P(\mathbf{X}) = \frac{e^{-H(\mathbf{X})}}{Z} = \frac{e^{-\sum_{i,j} J_{ij}x_i x_j - \sum_i \theta_i x_i}}{Z}$$

Inference tasks

- Computing marginal distribution $p(x_A)$ over a particular subset $A \subset V$ on nodes.
  - Computing conditional distribution $P(x_A|x_B)$
- Computing the most probable configurations. (MAP)

$$x = \underset{x \in \mathcal{X}^m}{\operatorname{argmax}} P(\mathbf{x})$$

Outline

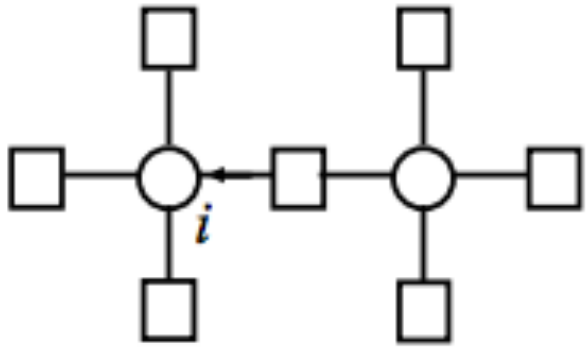
- Basics of graphical model
- Basics of message passing algorithm
- Variational free energy
  - Mean field approximation
  - TAP (Thouless, Anderson and Palmer)
  - Region Based approximation
    - Bethe free energy
    - Kikuchi approximation

Belief propagation

- BP is a method for computing marginal probability functions.
- The computed marginal probability is exact if the factor graph has no cycles.

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

$$m_{a\to i}(x_i) = \sum_{\mathbf{x}_a\setminus x_i} f_a(\mathbf{x_a}) \prod_{c\in N(i)\setminus a} m_{c\to i}(x_i)$$



- i is used as general index over variables, a over factors.

Belief propagation

In case this iteration converges, marginals are approximated by,

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

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

- In general LBP may not converge.
  - If it does, $b_i(x_i)$ may not be close to the true marginal $P(x_i)$.
  - The set of pseudomarginals $\mathbf{b}$ may not be realizable.

Outline

- Basics of graphical model
- Basics of message passing algorithm
- Variational free energy
  - Mean field approximation
  - TAP (Thouless, Anderson and Palmer)
  - Region Based approximation
    - Bethe free energy
    - Kikuchi approximation

Write down the energy function

Construct an approximation

Find the stationary condition

Variational free energy

- Variational method approximates an intractable distribution $P(\mathbf{X})$ of random variables $\mathbf{X} = (S_1, ..., S_N)$ by a tractable distribution $Q(\mathbf{X})$
- Q is chosen to minimize certain distance measure.

$$KL(Q||P) = \sum_{\mathbf{X}} Q(\mathbf{X}) \ln \frac{Q(\mathbf{X})}{P(\mathbf{X})} = \left\langle \ln \frac{Q}{P} \right\rangle_Q$$

where $\langle . \rangle_Q$ denotes the expectation with respect to Q

Variational free energy

To find the best approximate to $P = \frac{e^{-H(\mathbf{X})}}{Z}$

$$KL(Q||P) = \ln Z + E[Q] - S[Q]$$

where

- $S[Q] = -\sum_{\mathbf{X}} Q(\mathbf{X}) \ln Q(\mathbf{X})$ is the entropy of Q
- $E[Q] = \sum_{\mathbf{X}} Q(\mathbf{X}) H[\mathbf{X}]$ is called average energy

$$\Longrightarrow \min_Q KL(Q||P) = \ln Z + \min_Q \underbrace{(E[Q] - S[Q])}_{\text{Variational free energy}}$$

Variational free energy for Ising model

- The model under consideration is a Boltzmann machine.

$$P(\mathbf{X}) = \frac{e^{-H(\mathbf{X})}}{Z} = \frac{e^{-\sum_{i,j} J_{ij} x_i x_j - \sum_i \theta_i x_i}}{Z}$$

- For binary variable it is convenient to reparametrize these marginals as follows,

$$p_i(x_i = 1) = \frac{1 + m_i}{2}$$

Mean Field approximation

Find a factorized distribution that best describes the true distribution.

- For binary variable the most general factorized distribution has the form.

$$Q^{MF}(x) = \prod_i Q_i(x_i) = \prod_i \frac{(1 + x_i m_i)}{2}$$

- $KL(Q^{MF}||P) = E(Q^{MF}) - S(Q^{MF}) + \log(Z)$
- $E(Q^{MF}) = \sum Q^{MF} H(x) = -\sum_{ij} J_{ij} m_i m_j - \sum_i \theta_i m_i$
- $S(Q^{MF}) = -\sum_i Q^{MF} \ln Q^{MF} = -\sum_i \left(\frac{1+m_i}{2} \ln \frac{1+m_i}{2} + \frac{1-m_i}{2} \ln \frac{1-m_i}{2} \right)$

Mean Field approximation

How to solve?

$$\min_{m_i} KL(Q^{MF}||P)$$

- By taking derivative with respect to m_i
- $\frac{\partial}{\partial m_i}\left\{-\sum_{ij} J_{ij} m_i m_j - \sum_i \theta_i m_i + \sum_i \frac{1+m_i}{2} \ln \frac{1+m_i}{2} + \frac{1-m_i}{2} \ln \frac{1-m_i}{2} + \log(Z)\right\}$

Mean Field fixed points

$$\frac{\partial KL}{\partial m_i} = -\sum_{j \in N(i)} J_{ij} m_j - \theta_i + \log\left(\frac{m_i}{1 - m_i}\right)$$

- Fixed points of MF approximation:

$$m_i = \frac{\exp(\sum_j J_{ij} m_j + \theta_i) - \exp(-\sum_j J_{ij} m_j - \theta_i)}{\exp(\sum_j J_{ij} m_j + \theta_i) + \exp(-\sum_j J_{ij} m_j - \theta_i)}$$

$$\Rightarrow \boxed{m_i = \tanh(\sum_j J_{ij} m_j + \theta_i), \qquad i = 1, ..., N}$$

Mean Field

$$m_i = \tanh\Big(\sum_j J_{ij} m_j + \theta_i\Big), \qquad i = 1, \ldots, N \tag{1}$$

Note

- The intractable task of computing marginals has been replaced by the problem of solving a set of nonlinear equations.

Mean Field

$$m_i = \tanh(\sum_j J_{ij} m_j + \theta_i), \qquad i = 1, ..., N \tag{1}$$

Note

- The intractable task of computing marginals has been replaced by the problem of solving a set of nonlinear equations.
- These MF equations are run sequentially, i.e. we fix all m_j except m_i.

Mean Field

$$m_i = \tanh\Big(\sum_j J_{ij} m_j + \theta_i\Big), \qquad i = 1, \ldots, N \tag{1}$$

Note

- The intractable task of computing marginals has been replaced by the problem of solving a set of nonlinear equations.
- These MF equations are run sequentially, i.e. we fix all m_j except m_i.
- In each step MF free energy is convex. Equation (1) finds minimum in one step.

Mean Field

$$m_i = \tanh\Big(\sum_j J_{ij} m_j + \theta_i\Big), \qquad i = 1, \ldots, N \tag{1}$$

Note

- The intractable task of computing marginals has been replaced by the problem of solving a set of nonlinear equations.
- These MF equations are run sequentially, i.e. we fix all m_j except m_i.
- In each step MF free energy is convex. Equation (1) finds minimum in one step.
- This procedure can be interpreted as coordinate descent in the m_i

Mean Field

$$m_i = \tanh(\sum_j J_{ij} m_j + \theta_i), \qquad i = 1, ..., N \tag{1}$$

Note

- The intractable task of computing marginals has been replaced by the problem of solving a set of nonlinear equations.
- These MF equations are run sequentially, i.e. we fix all m_j except m_i.
- In each step MF free energy is convex. Equation (1) finds minimum in one step.
- This procedure can be interpreted as coordinate descent in the m_i
- Alternatively, all parameters m_i can be updated in parallel.

Mean Field

$$m_i = \tanh(\sum_j J_{ij} m_j + \theta_i), \qquad i = 1, ..., N \tag{1}$$

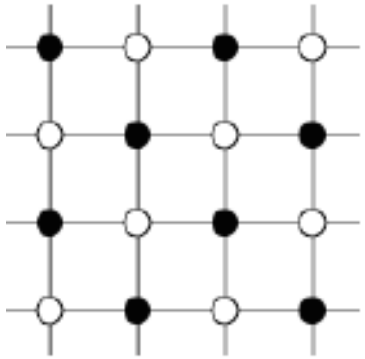
Note

- The intractable task of computing marginals has been replaced by the problem of solving a set of nonlinear equations.
- These MF equations are run sequentially, i.e. we fix all m_j except m_i.
- In each step MF free energy is convex. Equation (1) finds minimum in one step.
- This procedure can be interpreted as coordinate descent in the m_i
- Alternatively, all parameters m_i can be updated in parallel.
- Doesn't guarantee of decreasing the cost function at each iteration.
  - There might be many solutions to (1).
  - Some of the solutions may not be local minima

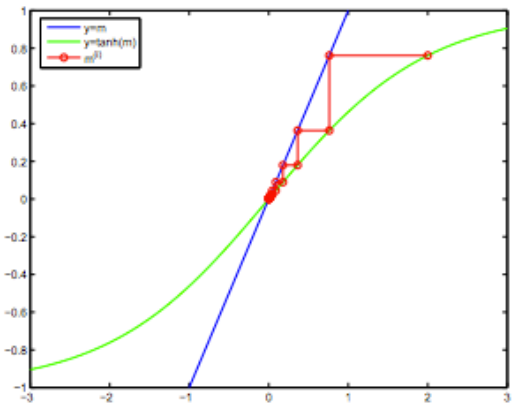
Mean Field

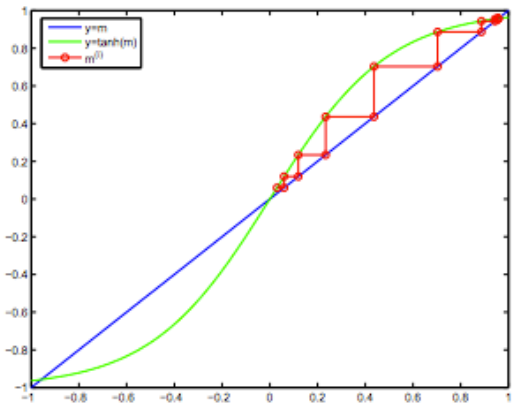
- In d-dimensional Ising model without the external magnetic field ($\theta = 0$) and having the same interaction $J_{ij} = \alpha$

$$m^{(t+1)} = \tanh(2d\alpha m^{(t)})$$



  - For $\alpha < \frac{1}{2d}$, the iteration converges to $\lim_{t\to\infty} m^{(t)} = 0$ (left figure)
  - For $\alpha > \frac{1}{2d}$, if $m^{(0)} \lesseqgtr 0 \Rightarrow \lim_{t\to\infty} m^{(t)} = \mp m^*$





[4]A. Montanari, Lecture notes for inference in graphical models,2011

Mean Field

- MF neglects the dependency between the random variables.

However,

- We get an upper bound on the exact free energy.

$$KL(Q^{MF}||P) = \underbrace{E(Q^{MF}) - S(Q^{MF})}_{=F[Q^{MF}] \text{ Variational MF energy}} - \underbrace{(-\log(Z))}_{\text{Exact free energy}}$$

Since $KL(Q^{MF}||P) \geq 0$

$$F(Q^{MF}) \geq -\log(Z)$$

Mean Field Method in general

- $P(x) = \frac{1}{Z} \prod_{a \in F} f_a(x_a)$ is True distribution
- $Q(x) = \prod_i q_i(x_i)$ is Approximate distribution

$$F^{MF}(Q) = \sum_i S(q_i) + \sum_{a \in F} \sum_{x_a} \prod_{x_i \in N(a)} q_i(x_i) \log f_a(x_a)$$

- We passed from $(|\mathcal{X}|^n - 1)$ to $n(|\mathcal{X}| - 1)$
- F^{MF} is no longer convex.

$$\min_Q F^{MF}(Q) \quad \text{subject to} \sum_{x_i} q_i(x_i) = 1$$

Mean Field Method in general

- Add Lagrange multiplier λ_i
- Find the stationary condition by $\frac{\partial L(Q,\lambda)}{\partial q_i(x_i)} = 0$

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

where

$$m_{a \to i}(x_i) = \exp\left(\sum_{x_j : j \in N(a) \setminus i} \log f_a(x_a) \prod_{j \in N(a) \setminus i} q_j(x_j) \right)$$

- A simple greedy algorithm for finding a stationary point consists in updating the q by iterating the above equations until convergence.

Outline

- Basics of graphical model
- Basics of message passing algorithm
- Variational free energy
  - Mean field approximation
  - TAP (Thouless, Anderson and Palmer)
  - Region Based approximation
    - Bethe free energy
    - Kikuchi approximation

The Legendre Transform and Plefka's Expansion

Plefka Expansion

- Don't restrict the approximate distribution Q to be product distributions
- Minimize free energy in two steps:
  - Constrained minimization in the family of distributions satisfying $\langle \mathbf{X} \rangle_Q = \mathbf{m}$ for fixed $\mathbf{m}$

$$G(\mathbf{m}) = \min_Q \{ F[Q] = E[Q] - S[Q] \ | \langle \mathbf{X} \rangle_Q = \mathbf{m} \}$$

  - Minimize $G(\mathbf{m})$ with respect to $\mathbf{m}$

Plefka Expansion

$$G(\mathbf{m}) = \min_{Q}\{F[Q] \mid \langle \mathbf{X} \rangle_Q = \mathbf{m}\}$$

By adding Lagrange multiplier λ
Then Lagrangian

$$G(\mathbf{m}, \lambda) = E[Q] - S[Q] - \sum_i \lambda_i(\langle x_i \rangle_Q - m_i)$$

$$G(\mathbf{m}, \lambda) = \sum_{\mathbf{X}} Q(\mathbf{X}) H[\mathbf{X}] - S[Q] - \sum_{\mathbf{x}} \sum_i \lambda_i x_i Q(\mathbf{X}) + \sum_i \lambda_i m_i$$

is the form of variational free energy, where $H[\mathbf{X}]$ is replaced by $H[\mathbf{X}] - \sum_i \lambda_i x_i$. We can construct such a gibbs free energy by adding a set of external auxiliary field.

$\Rightarrow Q_\lambda(\mathbf{X}) = \frac{1}{Z} e^{-H[\mathbf{X}] + \sum_i \lambda_i x_i}$

Plefka Expansion

The dual function is,

$$G(m_i) = \max_{\lambda_i} \left\{ \sum_i \lambda_i m_i - \log(Z(\lambda_i)) \right\}$$

- This equation known as Legendre transform between $\{\lambda_i\}$ and $\{m_i\}$.
- $Z(\lambda_i)$ is the normalizing constant for the Gibbs distribution

$$Q_\lambda(\mathbf{X}) = \frac{1}{Z_{\lambda_i}} e^{-H[\mathbf{X}] + \sum_i \lambda_i x_i} = \frac{1}{Z_{\lambda_i}} e^{-\sum_{i,j} J_{ij} x_i x_j - \sum_i \theta_i x_i + \sum_i \lambda_i x_i}$$

- Set $\theta \to 0$ by shifting the Lagrange multiplier $\lambda_i \to \lambda_i - \theta_i$
- $Z(\lambda_i) = \sum_{x_i} \exp(-\sum_{i,j} J_{ij} x_i x_j + \sum_i \lambda_i x_i)$

Plefka Expansion

$$G(m_i) = \max_{\lambda_i} \left\{ \sum_i \lambda_i m_i - \log\left(\sum_{x_i} \exp\left(-\sum_{i,j} \beta J_{ij} x_i x_j + \sum_i \lambda_i x_i\right)\right)\right\}$$

- Plefka expansion is derived by $J_{ij} \to \beta J_{ij}$, by Taylor expanding the Gibbs free energy around $\beta = 0$, where β is an inverse temperature in physics,

Notice

- For each term in Taylor expansion, one has to expand the Lagrange multiplier λ_i which maximize the Gibbs distribution as well as $\log(Z)$
- The auxiliary field is temperature dependent.

Plefka Expansion

- with $G_n = \frac{\partial^n}{\partial \beta^n} G(\mathbf{m})|_{\beta=0}$

$$G(\mathbf{m}) = G_0(\mathbf{m}) + \beta G_1(\mathbf{m}) + \frac{\beta^2}{2!} G_2(\mathbf{m}) + ...$$

Plefka Expansion

- with $G_n = \frac{\partial^n}{\partial \beta^n} G(\mathbf{m})|_{\beta=0}$

$$G(\mathbf{m}) = G_0(\mathbf{m}) + \beta G_1(\mathbf{m}) + \frac{\beta^2}{2!} G_2(\mathbf{m}) + \ldots$$

- $G_0(\mathbf{m}) = \sum_i \left\{ \frac{1+m_i}{2} \ln \frac{1+m_i}{2} + \frac{1-m_i}{2} \ln \frac{1-m_i}{2} \right\}$ Spins are entirely controlled by the auxiliary field.

Plefka Expansion

- with $G_n = \frac{\partial^n}{\partial \beta^n} G(\mathbf{m})|_{\beta=0}$

$$G(\mathbf{m}) = G_0(\mathbf{m}) + \beta G_1(\mathbf{m}) + \frac{\beta^2}{2!} G_2(\mathbf{m}) + \ldots$$

- $G_0(\mathbf{m}) = \sum_i \left\{ \frac{1+m_i}{2} \ln \frac{1+m_i}{2} + \frac{1-m_i}{2} \ln \frac{1-m_i}{2} \right\}$ Spins are entirely controlled by the auxiliary field.
- $G_1(\mathbf{m}) = -\sum_{i<j} J_{ij} m_i m_j$

Plefka Expansion

- with $G_n = \frac{\partial^n}{\partial \beta^n} G(\mathbf{m})|_{\beta=0}$

$$G(\mathbf{m}) = G_0(\mathbf{m}) + \beta G_1(\mathbf{m}) + \frac{\beta^2}{2!} G_2(\mathbf{m}) + \ldots$$

- $G_0(\mathbf{m}) = \sum_i \left\{ \frac{1+m_i}{2} \ln \frac{1+m_i}{2} + \frac{1-m_i}{2} \ln \frac{1-m_i}{2} \right\}$ Spins are entirely controlled by the auxiliary field.
- $G_1(\mathbf{m}) = -\sum_{i<j} J_{ij} m_i m_j$
- $G_2(\mathbf{m}) = -\frac{1}{2} \sum_{ij} J_{ij}^2 (1 - m_i^2)(1 - m_j^2)$
- ...

Plefka Expansion

with $G_n = \frac{\partial^n}{\partial \beta^n} G(\mathbf{m})|_{\beta=0}$

$$G(\mathbf{m}) = G_0(\mathbf{m}) + \beta G_1(\mathbf{m}) + \frac{\beta^2}{2!} G_2(\mathbf{m}) + ...$$

- $G_0 = \sum_i \left\{ \frac{1+m_i}{2} \ln \frac{1+m_i}{2} + \frac{1-m_i}{2} \ln \frac{1-m_i}{2} \right\} \Rightarrow$ MF variational entropy
- $G_1(\mathbf{m}) = -\sum_{i<j} J_{ij} m_i m_j \Rightarrow$ MF variational energy
- $G_2(\mathbf{m}) = -\frac{1}{2} \sum_{ij} J_{ij}^2 (1 - m_i^2)(1 - m_j^2)$
- ...$\Rightarrow$ Takes into account the higher order dependencies

TAP approximation

TAP approximation= Minimizing $G(m)$ for $\beta = 1$ and keeping only terms up to second order

$$G^{TAP}(m_i) = -\sum_{(ij)} J_{ij} m_i m_j + \sum_i \left\{ \frac{1+m_i}{2} \ln \frac{1+m_i}{2} + \frac{1-m_i}{2} \ln \frac{1-m_i}{2} \right\} \underbrace{- 1/2 \sum_{(ij)} J_{ij}^2 (1-m_i^2)(1-m_j^2)}_{\text{dependencies between rvs}}$$

- TAP takes in to account the dependencies between random variables.
- It's exact in the high temperature for certain classes of models (SK models).

TAP approximation

Fixed points of TAP approximation:

$$m_i = \tanh\left(\sum_{j \in N(i)} J_{ij} m_j + \frac{1}{2}(1 - 2m_i) \sum_{j \in N(i)} J_{ij}^2 m_j (1 - m_j) \right)$$

- Running these equations doesn't guarantee that TAP-Gibbs free energy decreases. (m_i appears on both sides)
- There is danger that radius of convergence (of taylor expansion) will be too small to obtain result for values of β we are interested in.

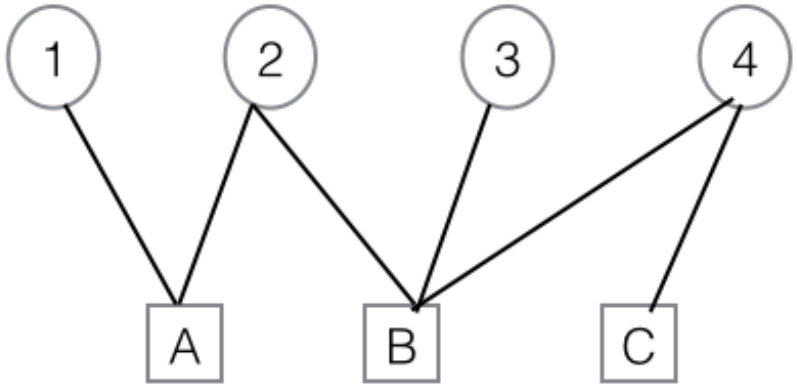
Outline

- Standard BP algorithm
- Junction tree algorithm
- Region Based free energy
  - Different types of region graph
- Special case: Bethe free energy
  - Stationary points of Bethe free energy = BP Fixed points
- Generalized belief propagation (GBP)
  - Stationary points of Region based free approximation

Outline

- Standard BP algorithm
- Junction tree algorithm
- Region Based free energy
  - Different types of region graph
- Special case: Bethe free energy
  - Stationary points of Bethe free energy = BP Fixed points
- Generalized belief propagation (GBP)
  - Stationary points of Region based free approximation

Message Passing - Computing the marginals

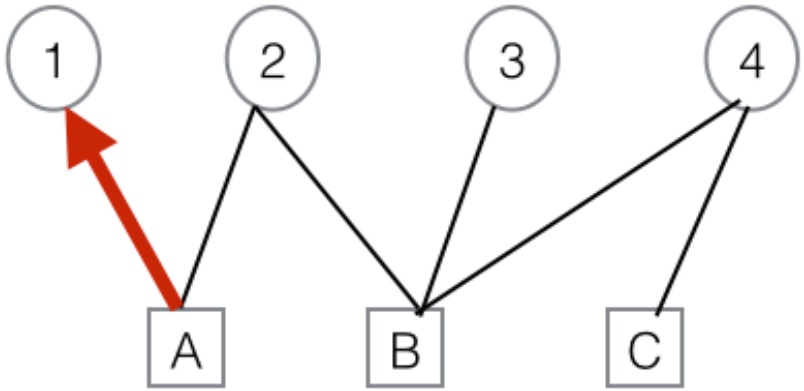


$$p(x_1, x_2, x_3, x_4) = f_A(x_1, x_2) f_B(x_2, x_3, x_4) f_C(x_4)$$

$$b_1(x_1) = p(x_1) = ?$$

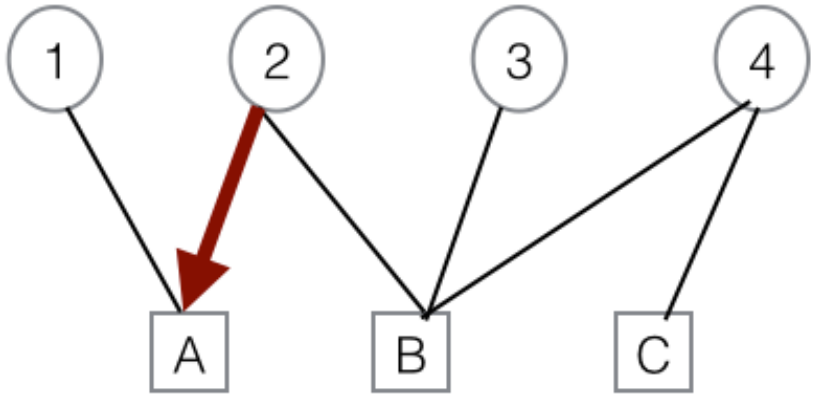
Message Passing

- $b_1(x_1) = m_{A \to 1}(x_1)$
-
-
-



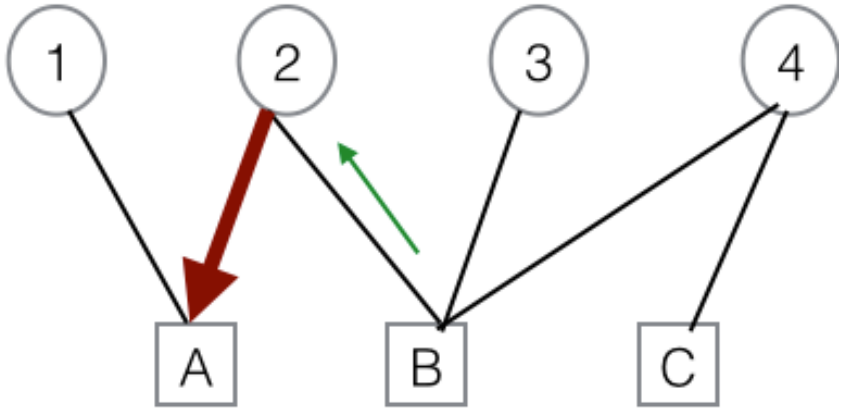
Message Passing

- $b_1(x_1) = m_{A \to 1}(x_1)$
- $b_1(x_1) = \sum_{x_2} f_A(x_1, x_2) m_{2 \to A}(x_2)$
-
-



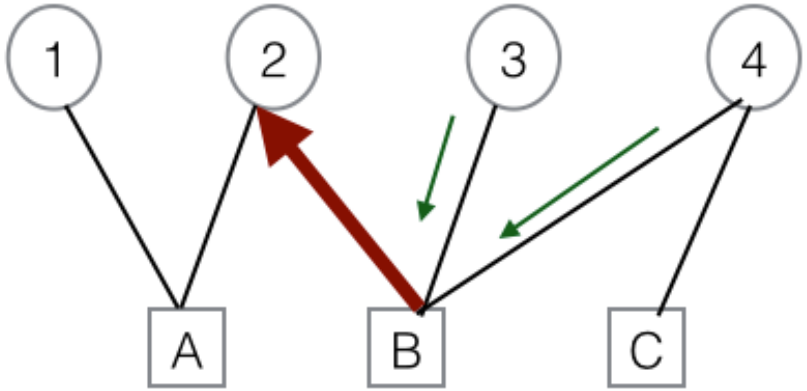
Message Passing

- $b_1(x_1) = m_{A \to 1}(x_1)$
- $b_1(x_1) = \sum_{x_2} f_A(x_1, x_2) m_{2 \to A}(x_2)$
- $b_1(x_1) = \sum_{x_2} f_A(x_1, x_2) m_{B \to 2}(x_2)$
-



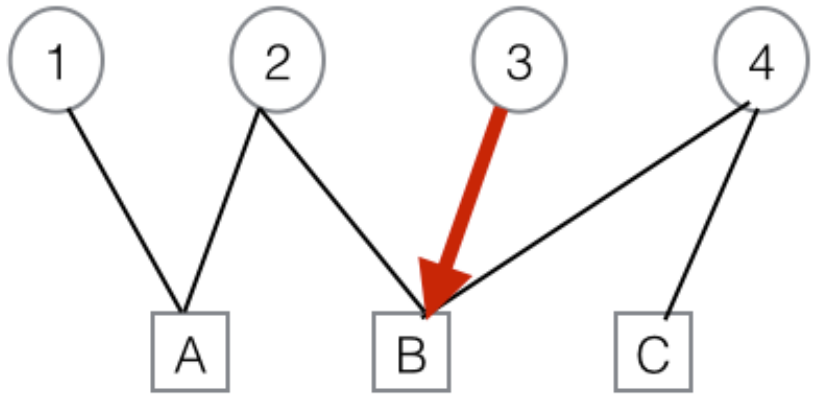
Message Passing

- $b_1(x_1) = m_{A\to 1}(x_1)$
- $b_1(x_1) = \sum_{x_2} f_A(x_1, x_2) m_{B\to 2}(x_2)$
- $b_1(x_1) = \sum_{x_2, x_3, x_4} f_A(x_1, x_2) f_B(x_2, x_3, x_4) m_{3\to B} m_{4\to B}$



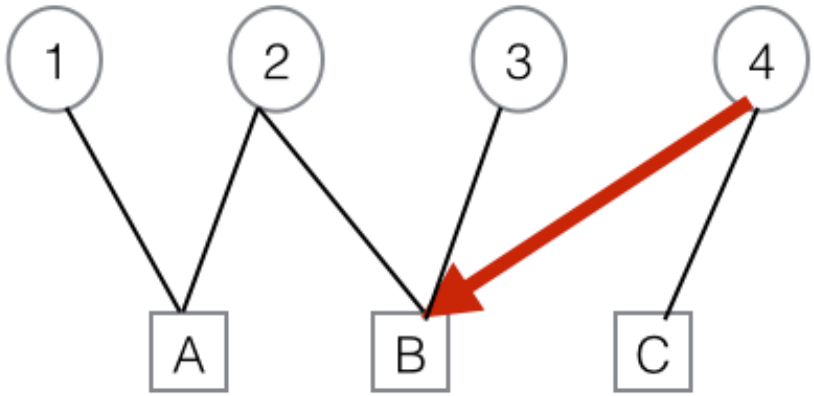
Message Passing

- $b_1(x_1) = m_{A\to 1}(x_1)$
- $b_1(x_1) = \sum_{x_2} f_A(x_1, x_2) m_{B\to 2}(x_2)$
- $b_1(x_1) = \sum_{x_2, x_3, x_4} f_A(x_1, x_2) f_B(x_2, x_3, x_4) m_{4\to B}(x_2)$



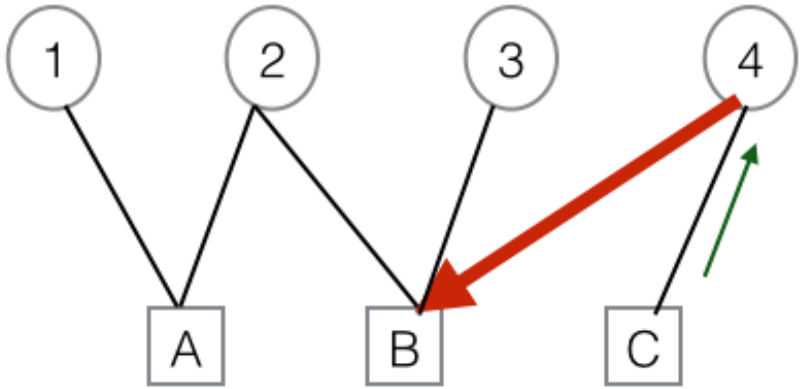
Message Passing

- $b_1(x_1) = m_{A\to 1}(x_1)$
- $b_1(x_1) = \sum_{x_2} f_A(x_1, x_2) m_{B\to 2}(x_2)$
- $b_1(x_1) = \sum_{x_2, x_3, x_4} f_A(x_1, x_2) f_B(x_2, x_3, x_4) m_{4\to B}(x_2)$



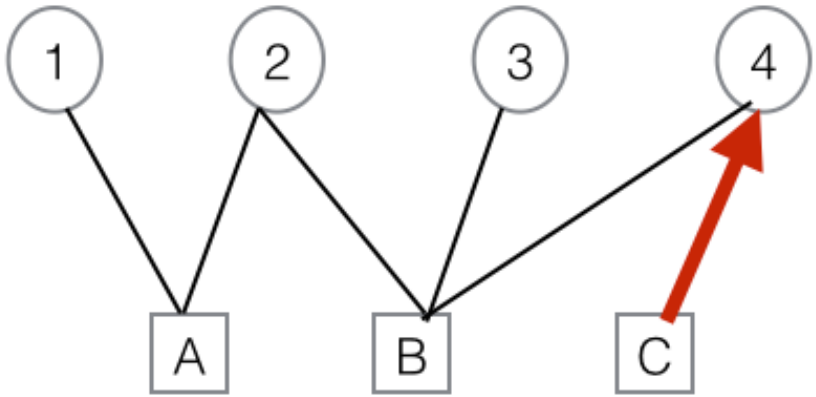
Message Passing

- $b_1(x_1) = m_{A\to 1}(x_1)$
- $b_1(x_1) = \sum_{x_2} f_A(x_1, x_2) m_{B\to 2}(x_2)$
- $b_1(x_1) = \sum_{x_2, x_3, x_4} f_A(x_1, x_2) f_B(x_2, x_3, x_4) m_{C\to 4}(x_4)$



Message Passing

- $b_1(x_1) = m_{A\to 1}(x_1)$
- $b_1(x_1) = \sum_{x_2} f_A(x_1, x_2) m_{B\to 2}(x_2)$
- $b_1(x_1) = \sum_{x_2, x_3, x_4} f_A(x_1, x_2) f_B(x_2, x_3, x_4) f_C(x_4)$



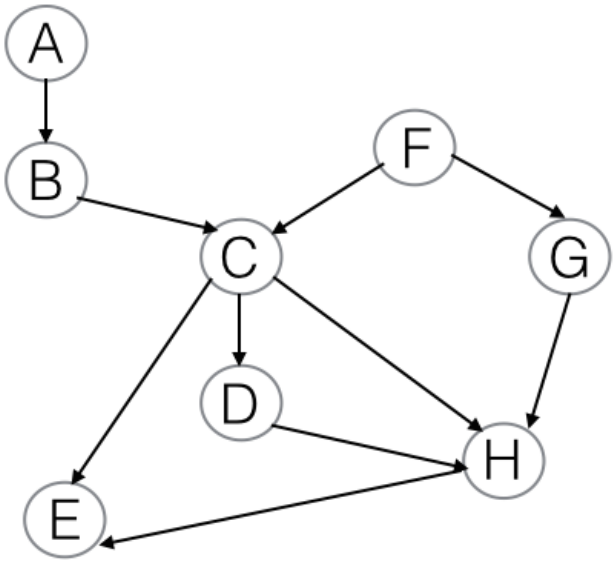
Outline

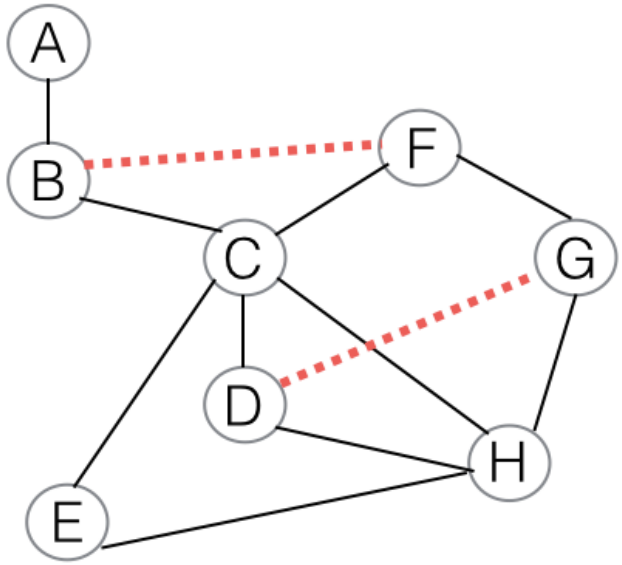
- Standard BP algorithm
- Junction tree algorithm
- Region Based free energy
  - Different types of region graph
- Special case: Bethe free energy
  - Stationary points of Bethe free energy = BP Fixed points
- Generalized belief propagation (GBP)
  - Stationary points of Region based free approximation

Junction Tree algorithm

- Works for general graph
  - Tree shape graphs
  - Graphs with cycles
  - Directed graphs
  - Undirected graphs
- Remove cycles by clustering nodes into cliques.
- Perform Belief Propagation on cliques.
- Exact inference of (clique) marginals.

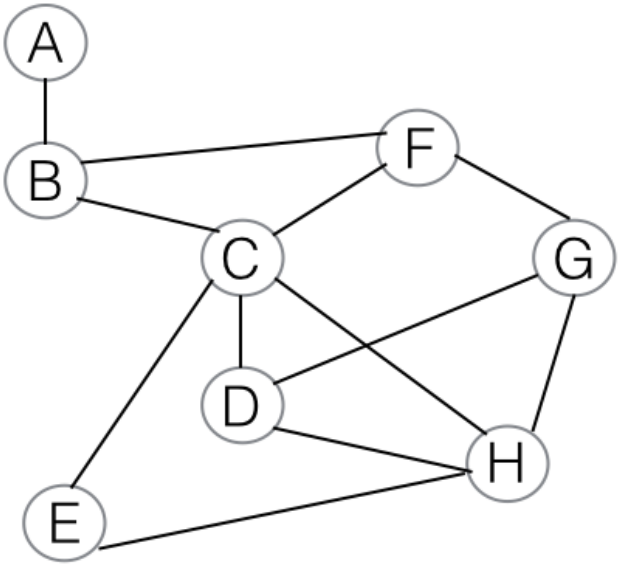
Junction Tree algorithm - Moralization

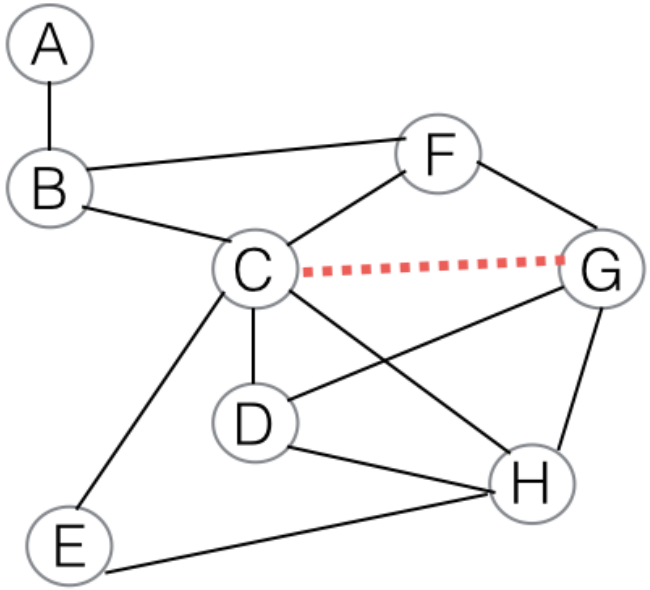




- we first moralize the graph by connecting all unconnected parents. After this we make the graph an undirected graph

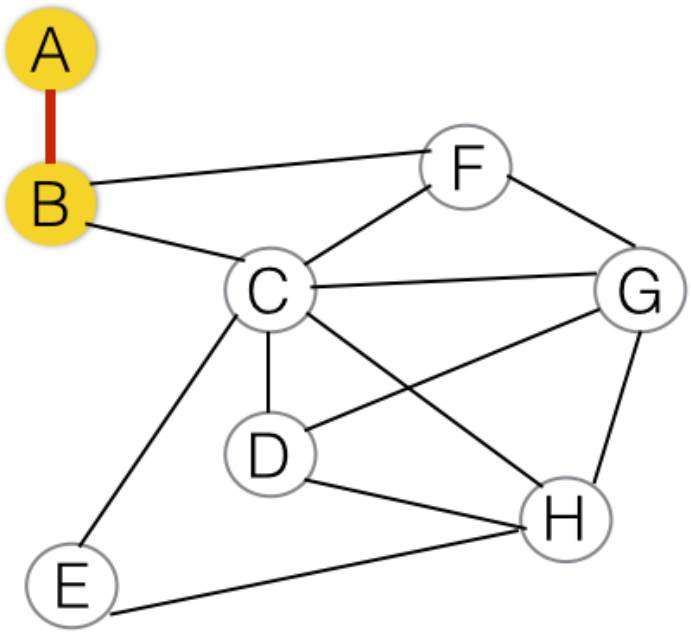
Junction Tree algorithm- Triangulation



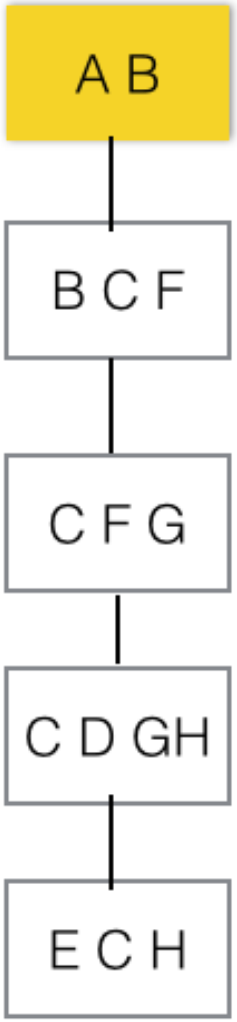


- Triangulation i.e. for any given cycle there is an edge between any two non-successive nodes in the cycle

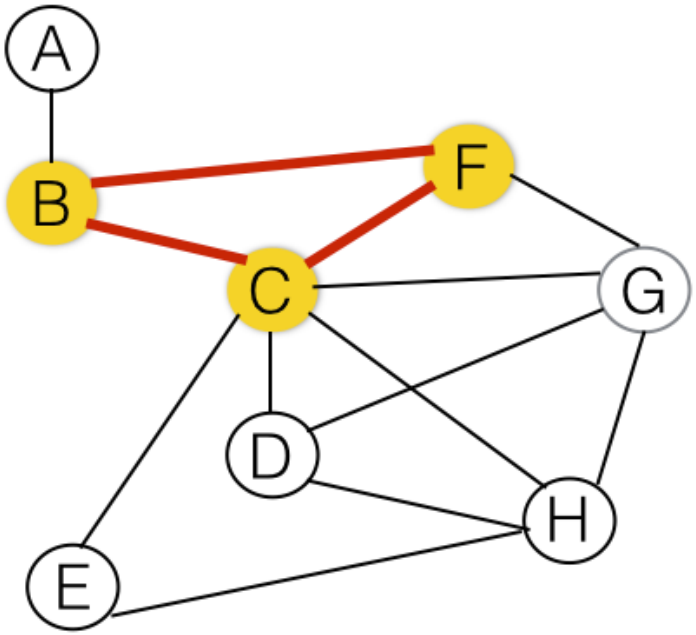
Junction Tree algorithm

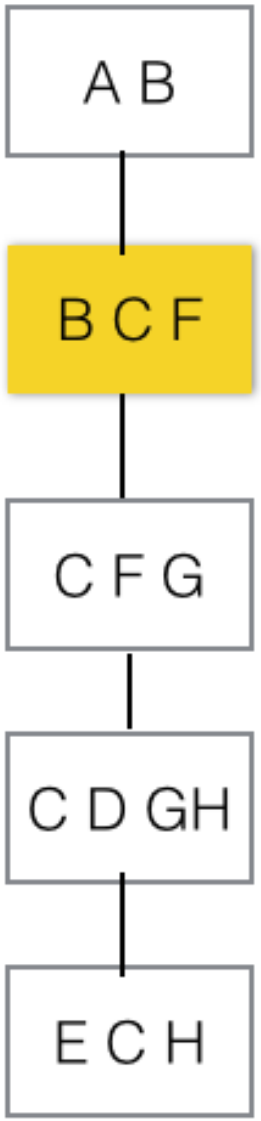


$$\psi_{C_1}(x_A, x_B) = \psi_{A,B}(x_A, x_B)$$



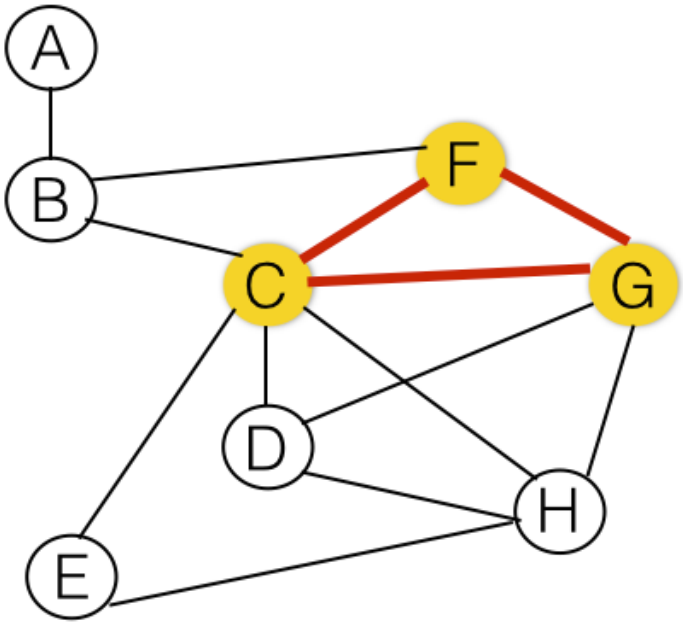
Junction Tree algorithm





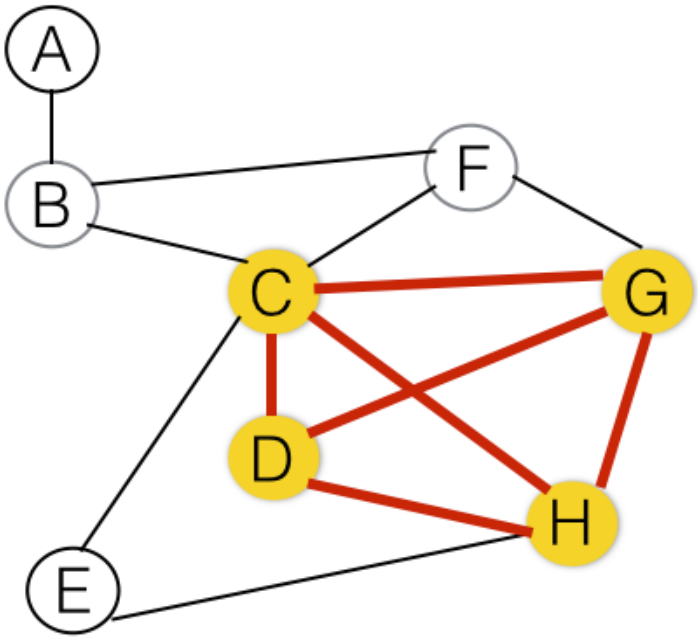
$$\psi_{C_2}(x_B, x_C, x_F) = \psi_{B,C}(x_B, x_C)\psi_{C,F}(x_C, x_F)$$

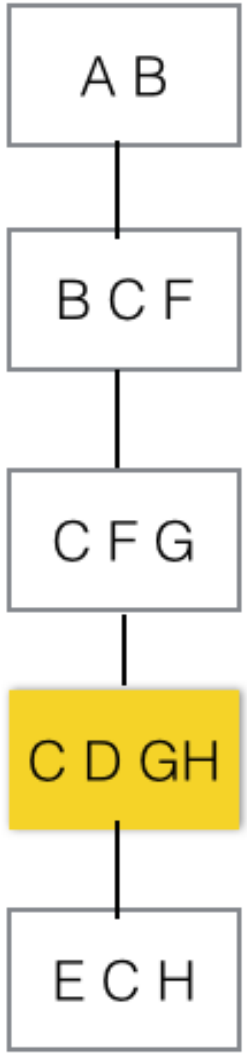
Junction Tree algorithm



$$\psi_{C_3}(x_C, x_F, x_G) = \psi_{C,F}(x_C, x_F)\psi_{F,G}(x_F, x_G)$$

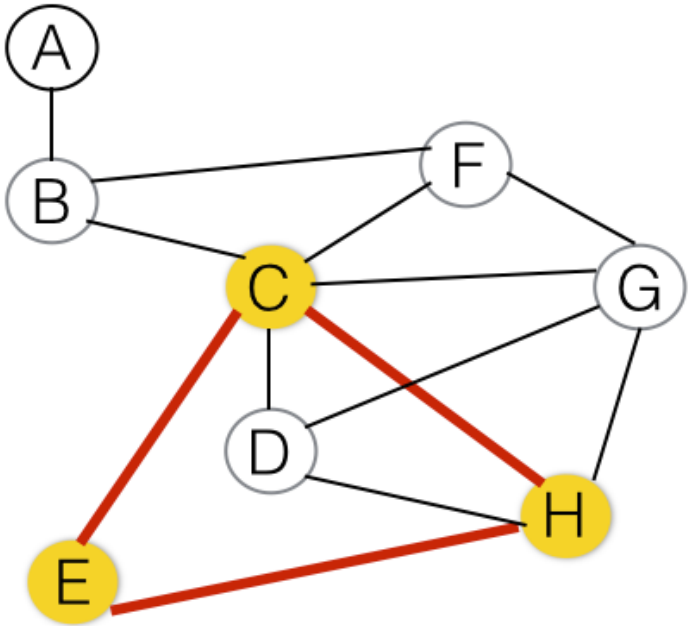
Junction Tree algorithm

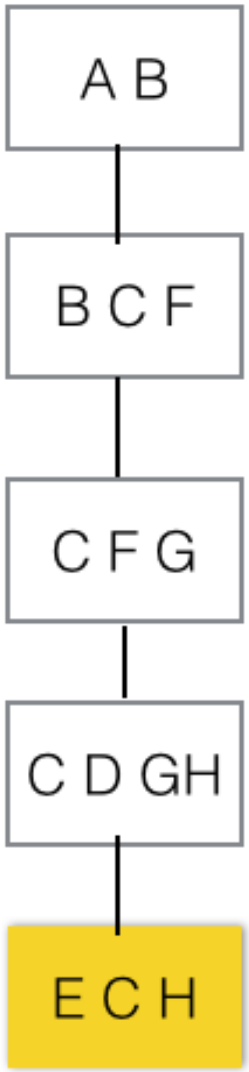




$$\psi_{C_4}(x_C, x_D, x_G, x_H) =$$
$$\psi_{C,D,H}(x_C, x_D, x_H)\psi_{D,G,H}(x_D, x_G, x_H)$$

Junction Tree algorithm

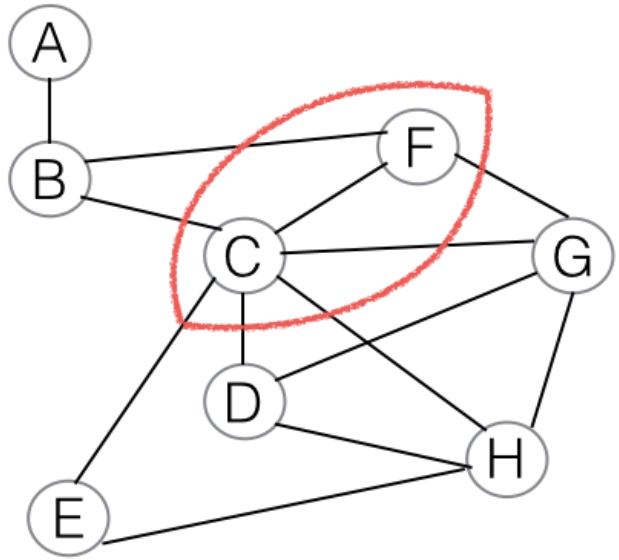


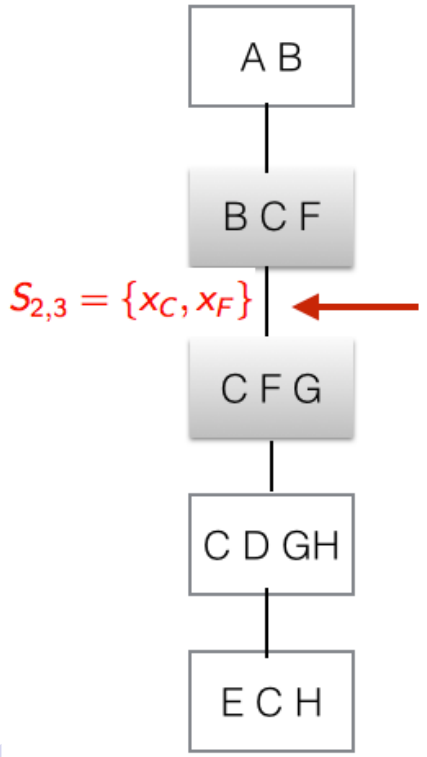


$$\psi_{C_5}(x_C, x_E, x_H) = \psi_{C,E,H}(x_C, x_E, x_H)$$

Independence in junction tree

- suppose
  - T is a junction tree for graph G.
  - Consider cliques C_i and C_j with separator $S_{ij} = C_i \cap C_j$
  - Variables X and Y are on opposite site of separator.
- X and Y are independent given S_{ij}





Junction Tree algorithm

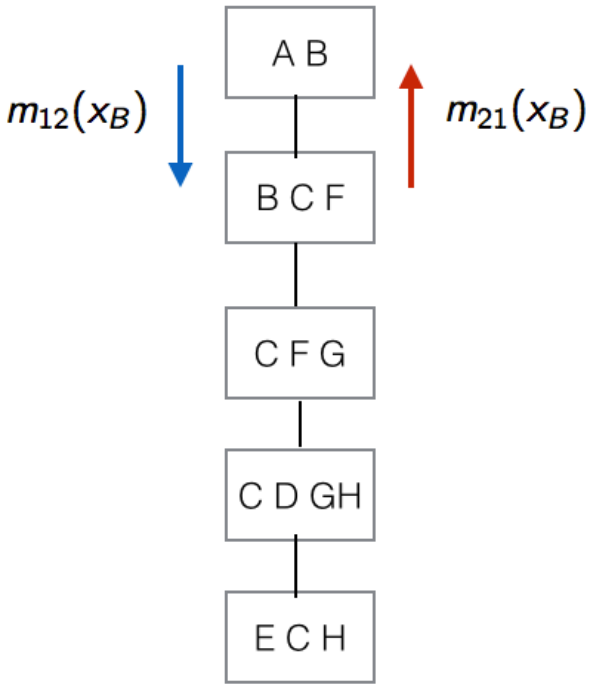
Given junction tree and potentials on the cliques, the messages from clique C_i to C_j is

$$m_{ij}(x_{S_{ij}}) = \sum_{C_i \setminus S_{ij}} \psi_{C_i}(x_{C_i}) \prod_{k \in N(i) \setminus j} m_{ki}(x_{S_{ki}})$$

- S_{ij}: nodes shared by i and j
- $N(i)$: neighboring cliques of i
- The marginal distribution of any cliques are

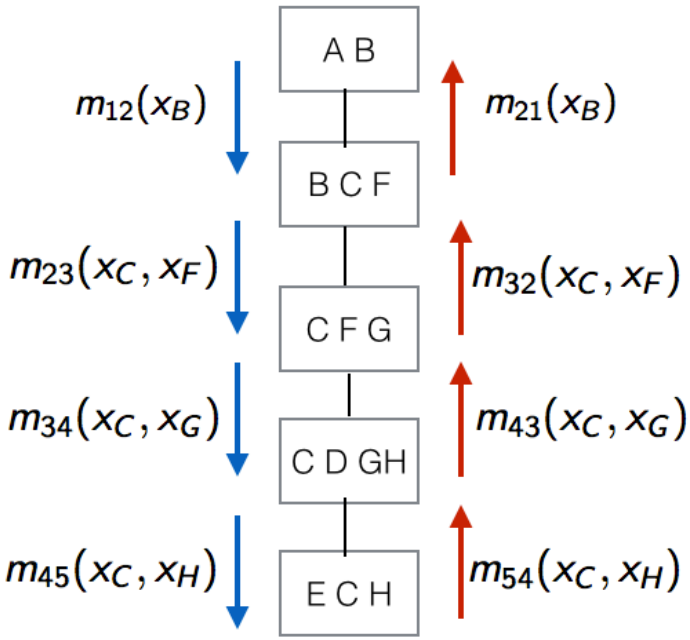
$$p(x_{C_i}) = \psi_{C_i} \prod_{k \in N(i)} m_{ki}(x_{S_{ki}})$$

$$p(x_{S_{ij}}) = m_{ij} m_{ji}$$



Junction Tree algorithm

- $m_{12}(x_B) = \sum_{x_A} \psi_{C_1}(x_A, x_B)$
- $m_{23}(x_C, x_F) = \sum_{x_B} \psi_{C_2}(x_B, x_C, x_F) m_{12}(x_B)$
- $m_{34}(x_C, x_G) = \sum_{x_F} \psi_{C_3}(x_C, x_F, x_G) m_{23}(x_C, x_F)$
- $m_{45}(x_C, x_H) = \sum_{x_D, x_G} \psi_{C_4}(x_C, x_D, x_G, x_H) m_{34}(x_C, x_G)$



Outline

- Standard BP algorithm
- Junction tree algorithm
- Region Based free energy
  - Different types of region graph
- Special case: Bethe free energy
  - Stationary points of Bethe free energy = BP Fixed points
- Generalized belief propagation (GBP)
  - Stationary points of Region based free approximation

Variational free energy

To find the best approximate to $P = \frac{1}{Z} \prod_{c \in cliques} \phi_c(x_c)$

$$KL(Q||P) = \sum_{\mathbf{X}} Q(x) \ln Q(x) - \sum_{x} Q(x) \ln p(x)$$

where

- $U[Q] = -\sum_{x} Q(x) \ln Q(x)$ is the entropy of Q
- $H[Q] = -\sum_{c \in cliques} \sum_{x_c} Q(x_c) \log \phi_c(x_c)$ is called average energy

$$\Longrightarrow \min_{Q} KL(Q||P) = \ln Z + \min_{Q} \underbrace{(U[Q] - H[Q])}_{\text{Variational free energy}}$$

Variational Free energy

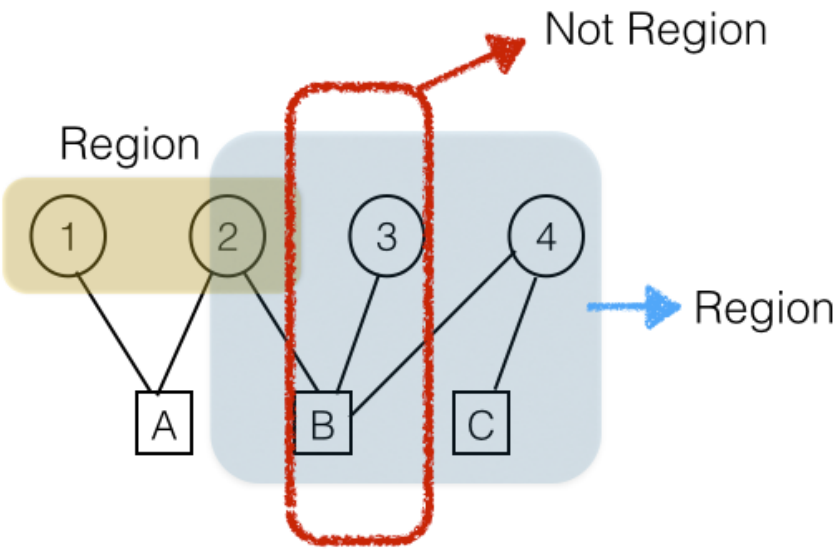
- Two solution methods to

$$\min_{Q} F[Q]$$

  - Approximate F[Q]
    - Region Based approximation $\Longrightarrow F_{\mathcal{R}}(q_{\mathcal{R}})$
  - Choose a simpler form of Q
    - Mean Field Approximation $\Longrightarrow Q = \prod q_i$

Region Based free energy

- We decompose the system into subsystems and then approximate the free energy by combining the free energies of the subsystems
- Group nodes in to (possibly overlapping) clusters.
- In each region, all variable nodes connected to any included factor nodes are included.



- The sets of nodes $\{1, 2\}$, $\{B, C, 2, 3, 4\}$ could be regions.
- $\{B, 3\}$ could not be a region.

Region Based free energy

- The overall energy is the sum of the free energies of all the regions.

Region Based free energy

- The overall energy is the sum of the free energies of all the regions.
- If some of the large regions overlap, subtract out the free energies of these overlap region.

Region Based free energy

- The overall energy is the sum of the free energies of all the regions.
- If some of the large regions overlap, subtract out the free energies of these overlap region.
- Each factor and variable node should be counted exactly once.

Region Based free energy

- The overall energy is the sum of the free energies of all the regions.
- If some of the large regions overlap, subtract out the free energies of these overlap region.
- Each factor and variable node should be counted exactly once.
- For every factor node a and every variable node i in a set of regions $\mathcal{R}$, the counting number is

$$\sum_{R \in \mathcal{R}} c_R \mathrm{I}(a \in F_R) = \sum_{R \in \mathcal{R}} c_R \mathrm{I}(i \in V_R) = 1$$

where $\mathrm{I}(x \in \mathcal{S}) = 1$ if $x \in \mathcal{S}$

Region Based free energy

- Region base free energy for a set of region $\mathcal{R}$ is

$$F_{\mathcal{R}}(b_R) = U_{\mathcal{R}}(b_R) - H_{\mathcal{R}}(b_R)$$

- Count every node once.
  - $U_{\mathcal{R}}(b_R) = \sum_{R \in \mathcal{R}} c_R U_R(b_R) \Longrightarrow$ region based average energy
  - $H_{\mathcal{R}}(b_R) = \sum_{R \in \mathcal{R}} c_R H_R(b_R) \Longrightarrow$ region based approximate entropy

Region Based free energy

if

$$\sum_{R \in \mathcal{R}} c_R \mathrm{I}(i \in F_R) = 1 \text{for all } a \in F$$

and

$$b_R(\mathbf{x}_R) = p_R(\mathbf{x}_R)$$

$\Longrightarrow$ The average energy becomes exact.

$$U_{\mathcal{R}}(b_R) = \sum_{R \in \mathcal{R}} c_R U_R(b_R) = -\sum_{R \in \mathcal{R}} c_R \sum_{\mathbf{x}_R} b_R(\mathbf{x}_R) \sum_{a \in F_R} \ln f_a(\mathbf{x}_a)$$

Exact energy $\Rightarrow U = \sum_{x \in S} p(\mathbf{x}) E(x) = -\sum_a \sum_{\mathbf{x}_a} p_a(\mathbf{x}_a) \ln f_a(\mathbf{x}_a)$

Region Based free energy

- Counting each variable node and factor node exactly once, results in exactness of the average energy.
- However, the region based entropy is still an approximation.

$$H_{\mathcal{R}}(b_R) = \sum_{R \in \mathcal{R}} c_R H_R(b_R) = -\sum_{R \in \mathcal{R}} c_R \sum_{\mathbf{x}_R} b_R(\mathbf{x}_R) \ln b_R(\mathbf{x}_R)$$

- We are interested in the accuracy of $H_{\mathcal{R}}(b_R)$ near its maximum.

$$\min_{b_R} F_{\mathcal{R}}(b_R) = \min_{b_R} \{U_{\mathcal{R}}(b_R) - H_{\mathcal{R}}(b_R)\}$$

- $H_{\mathcal{R}}(b_R)$ should achieve its maximum when all beliefs $b_R(\mathbf{x}_R)$ are uniform. (Maxent normal)

Outline

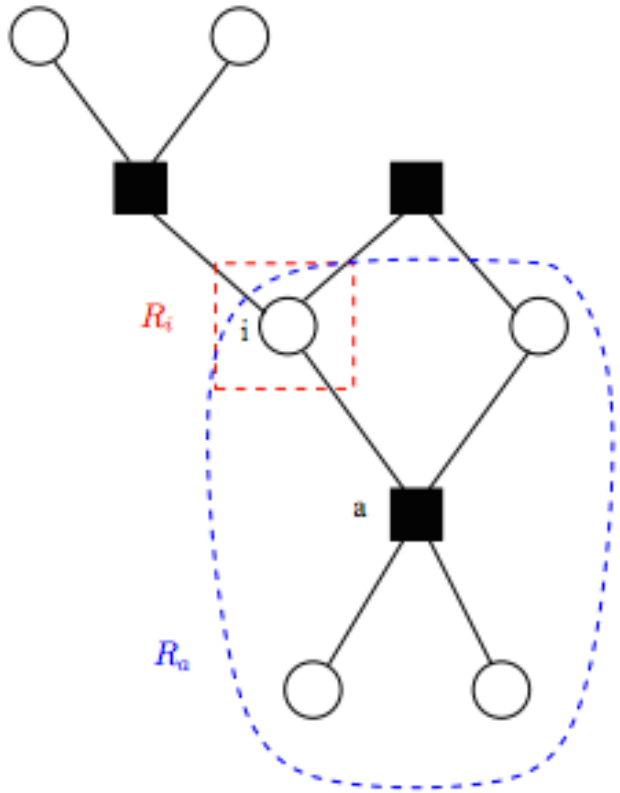
- Standard BP algorithm
- Junction tree algorithm
- Region Based free energy
  - Different types of region graph
- Special case: Bethe free energy
  - Stationary points of Bethe free energy = BP Fixed points
- Generalized belief propagation (GBP)
  - Stationary points of Region based free approximation

Bethe Free energy

Regions are $\mathcal{R} = \{R_i, R_a, i \in V, a \in F\}$

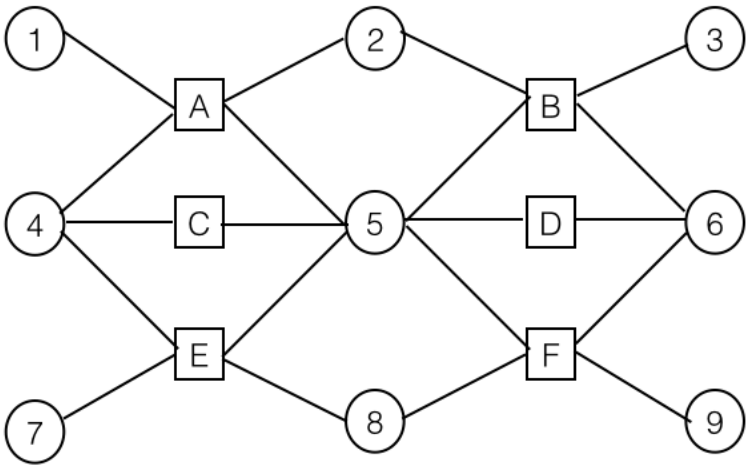
- $R_i = (\{i\}, 0, 0)$
- $R_a = (\{\mathcal{N}(a)\}, \{a\}, \{(i, a) : i \in \mathcal{N}(a)\})$

- Large regions containing a single factor node a and all attached variable nodes. $c_r = 1$
- Small regions containing a single variable node $c_r = 1 - d_i$ where $d_i = |\mathcal{N}(i)|$
- R_1 is subregion of R_2 if $R_1 \subset R_2$



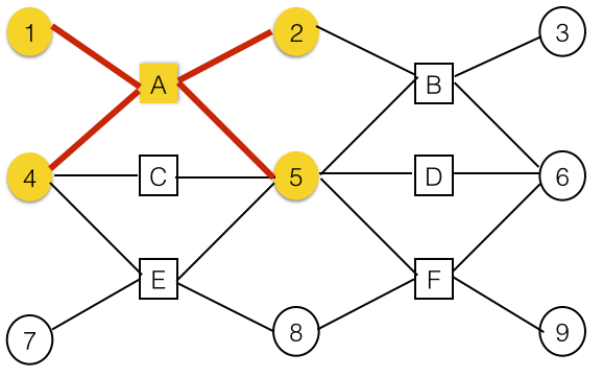
Bethe Free energy

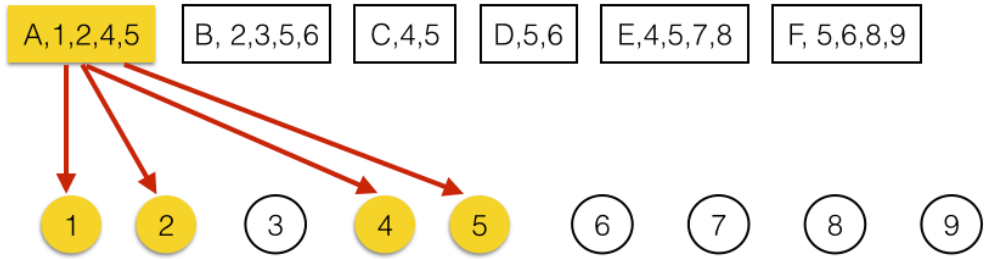
- Bethe region graph for the following factor graph



Bethe Free energy

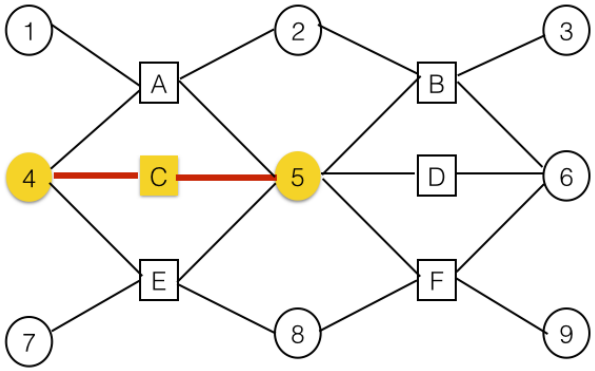
- Bethe region graph for the following factor graph

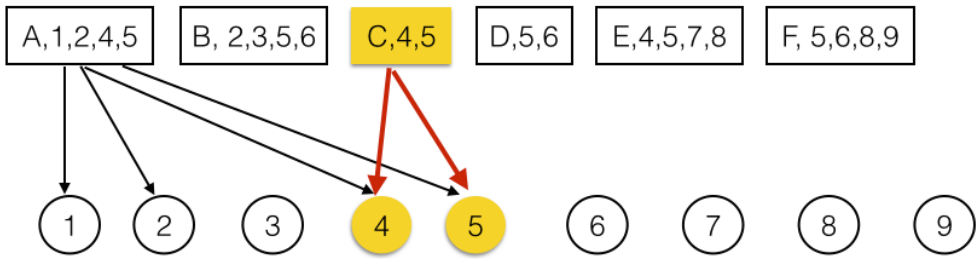




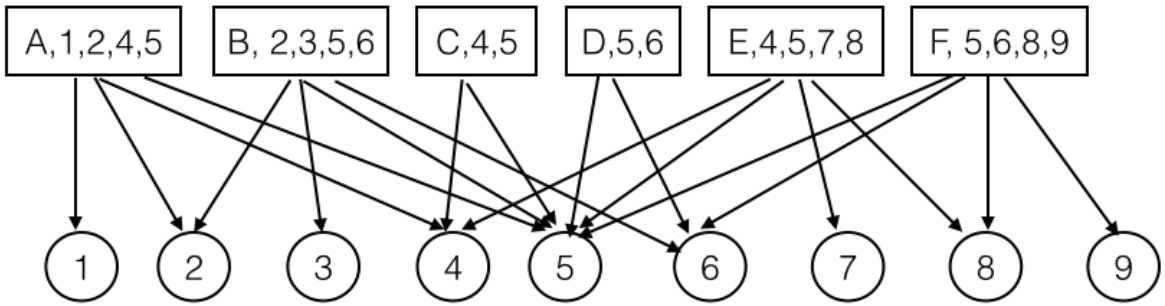
Bethe Free energy

- Bethe region graph for the following factor graph





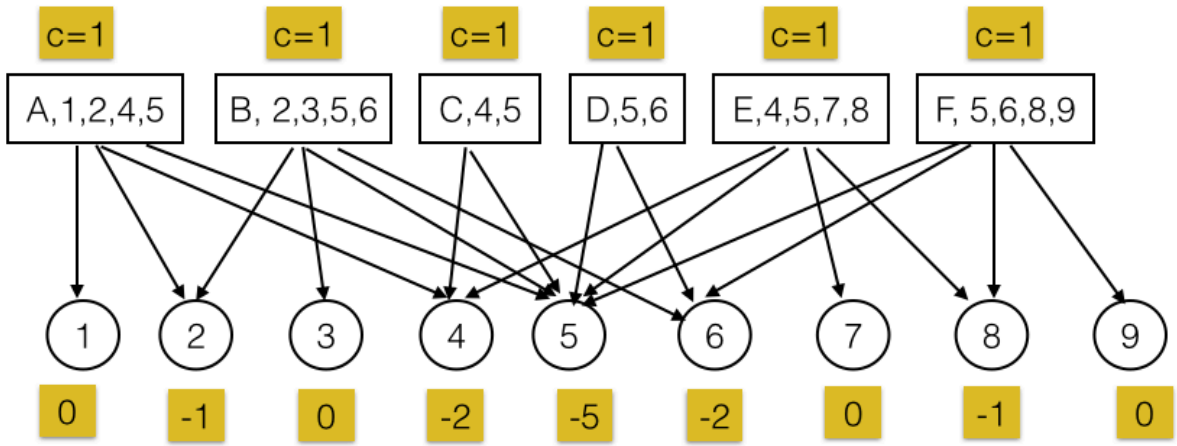
Bethe Free energy



$c_r = 1$ for $r \in R_a$
$c_r = 1 - d_i$ for $r \in R_i$

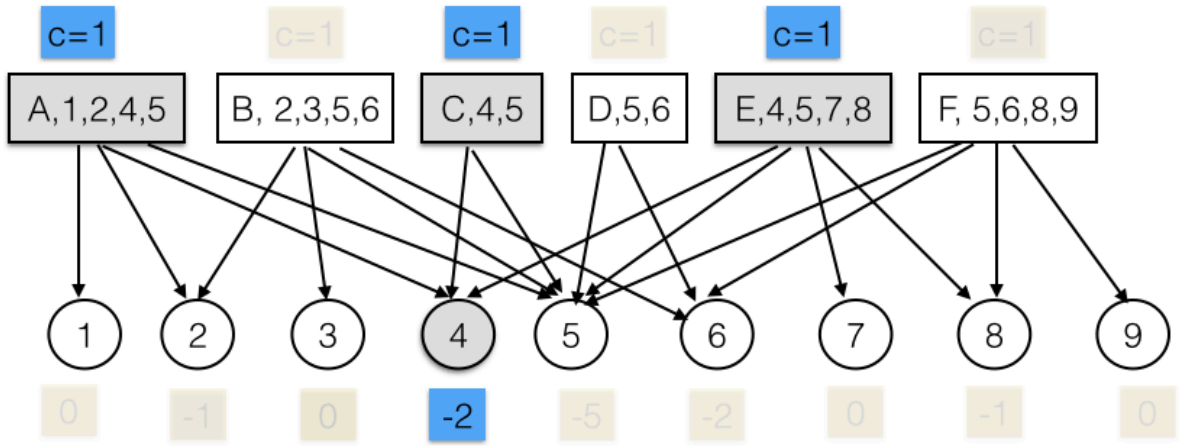
Bethe Free energy

- Assigning counting number to the regions.



Bethe Free energy

- Every variable node and factor node is counted once.



Bethe Free energy

- Bethe free energy:

$$F_{Bethe} = U_{Bethe} - H_{Bethe}$$

- Bethe average energy:

$$U_{Bethe} = -\sum_{a}\sum_{x_a} b_a(x_a) \ln f_a(x_a)$$

- Bethe entropy:

$$H_{Bethe} = -\sum_{a}\sum_{x_a} b_a(x_a) \ln b_a(x_a) + \sum_{i}(d_i - 1)\sum_{x_i} b_i(x_i) \ln b_i(x_i)$$

Bethe Free energy - Maxent normal

- Global maximum of Bethe entropy is achieved when the beliefs $b_i(x_i)$, $b_a(x_a)$ are uniform.

$$H_{Bethe} = \sum_i H(b_i) - \sum_a I(b_a)$$

where

$$H(b_i) = -\sum_{x_a} b_i(x_i) \ln b_i(x_i)$$

$$I(b_a) = -\Big(\sum_{x_a} b_a(x_a) \ln b_a(x_a) - \sum_{i \in N(a)} H(b_i) \Big)$$

- Maximum of $H(b_i)$ achieved when $b_i(x_i)$ has uniform dist.
- $I(b_a) \geq 0 \rightarrow$ when the beliefs are uniform, $I(b_a) = 0$

Constrained Bethe free energy

Constrained Bethe free energy enforces the beliefs to obey:

- The normalization constrains:

$$\sum_{x_i} b_i(x_i) = 1$$

$$\sum_{x_a} b_a(x_a) = 1$$

- Consistency constraints

$$\sum_{x_a \setminus x_i} b_a(x_a) = b_i(x_i)$$

- Inactive Constraint $\Rightarrow$ Complementary slackness

$$0 \leq b_i(x_i) \leq 1$$

$$0 \leq b_a(x_a) \leq 1$$

Minimizing Constrained Bethe free energy

Theorem:
Stationary points of the constrained Bethe free energy are BP fixed points.

$$
\begin{aligned}
\underset{b}{\text{minimize}} \quad & F_{Bethe} \\
\text{subject to} \quad & \sum_{x_i} b_i(x_i) = 1 \\
& \sum_{x_a} b_a(x_a) = 1 \\
& \sum_{x_a \setminus x_i} b_a(x_a) = b_i(x_i) \\
& b_a(x_a), b_i(x_i) \geq 0
\end{aligned}
$$

Minimizing Constrained Bethe free energy

- Lagrangian:

$$L = F_{Bethe} + \sum_i \gamma_i \left\{ \sum_{x_i} b_i(x_i) - 1 \right\} + \sum_a \sum_{i \in N(a)} \sum_{x_i} \lambda_{ai}(x_i) \left\{ \sum_{x_a \setminus x_i} b_a(x_a) - b_i(x_i) \right\}$$

Minimizing Constrained Bethe free energy

- Lagrangian:

$$L = F_{Bethe} + \sum_i \gamma_i \left\{ \sum_{x_i} b_i(x_i) - 1 \right\} + \sum_a \sum_{i \in N(a)} \sum_{x_i} \lambda_{ai}(x_i) \left\{ \sum_{x_a \setminus x_i} b_a(x_a) - b_i(x_i) \right\}$$

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

Minimizing Constrained Bethe free energy

- Lagrangian:

$$L = F_{Bethe} + \sum_i \gamma_i \left\{ \sum_{x_i} b_i(x_i) - 1 \right\} + \sum_a \sum_{i \in N(a)} \sum_{x_i} \lambda_{ai}(x_i) \left\{ \sum_{x_a \setminus x_i} b_a(x_a) - b_i(x_i) \right\}$$

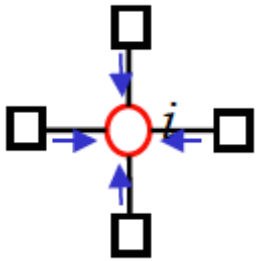
- $\frac{\partial L}{\partial b_i(x_i)} = 0 \Longrightarrow b_i(x_i) = \exp\left(\frac{1}{d_i - 1} \{ 1 - \gamma_i + \sum_{a \in N(i)} \lambda_{ai}(x_i) \} \right)$
- $\frac{\partial L}{\partial b_a(x_a)} = 0 \Longrightarrow b_a(x_a) = \exp\left(-E_a(x_a) + \sum_{a \in N(i)} \lambda_{ai}(x_i) \right)$

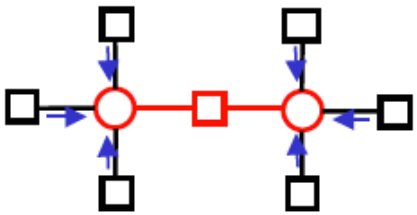
Bethe Fixed points

Define

$$\lambda_{ai}(x_i) = \ln \prod_{b \in N(i) \setminus a} m_{b \to i}(x_i)$$

Obtain BP equations:



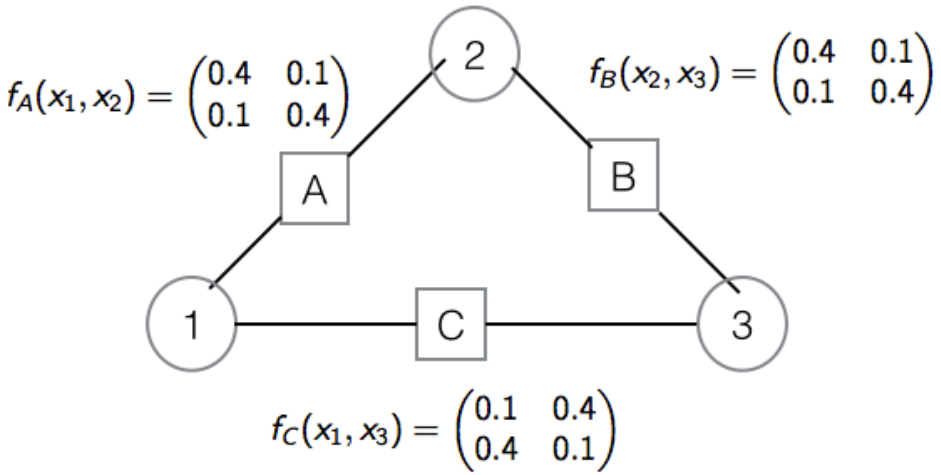


$$b_i(x_i) \propto \prod_{a \in N(i)} m_{a \to i}(x_i)$$

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

Unrealizable beliefs

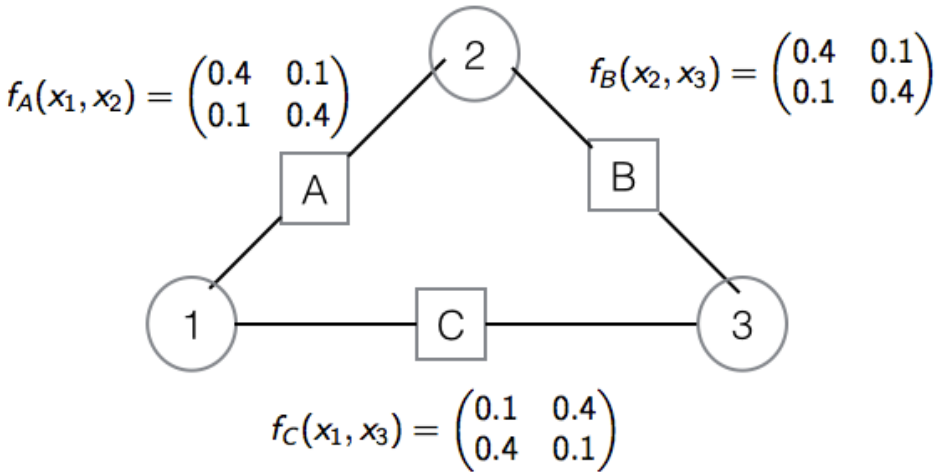
- $b_A(x_1, x_2) = \begin{pmatrix} 0.4 & 0.1 \\ 0.1 & 0.4 \end{pmatrix}$
- $b_B(x_2, x_3) = \begin{pmatrix} 0.4 & 0.1 \\ 0.1 & 0.4 \end{pmatrix}$
- $b_C(x_1, x_3) = \begin{pmatrix} 0.1 & 0.4 \\ 0.4 & 0.1 \end{pmatrix}$



- $b_1(x_1) = b_2(x_2) = b_3(x_3) = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$

Unrealizable beliefs

- $b_A(x_1, x_2) = \begin{pmatrix} 0.4 & 0.1 \\ 0.1 & 0.4 \end{pmatrix}$
- $b_B(x_2, x_3) = \begin{pmatrix} 0.4 & 0.1 \\ 0.1 & 0.4 \end{pmatrix}$
- $b_C(x_1, x_3) = \begin{pmatrix} 0.1 & 0.4 \\ 0.4 & 0.1 \end{pmatrix}$



- $b_1(x_1) = b_2(x_2) = b_3(x_3) = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$
- There is no $b(x_1, x_2, x_3)$!

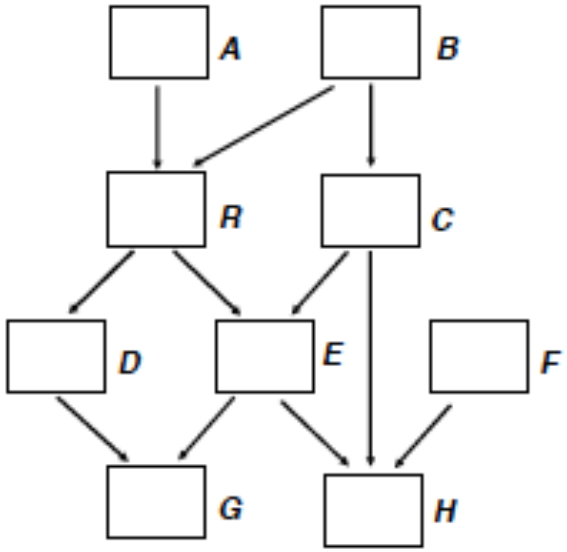
Region based energy

- How to select a set of regions $\mathcal{R}$ and and counting number c_R?
- Some methods are:
  - Bethe method
  - Junction Graph method
  - Cluster variation method
  - Region Graph method

Region Graph

- Region graph is a directed acyclic graph, $R \rightarrow R' \Rightarrow R' \subseteq R$.
- If there is a directed path between R and R', we say R is ancestor of R' , $R \in A(R')$ and R' is a descendant of R, $R' \in D(R)$
- In in a region graph these set of conditions satisfied,

$$c_R = 1 - \sum_{R' \in A(R)} c'_R \text{ for all } R \in \mathcal{R}$$



Region Graph Condition

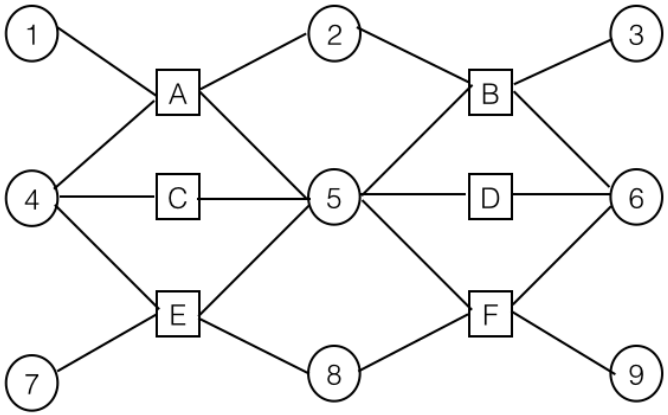
- Every nodes is counted once:

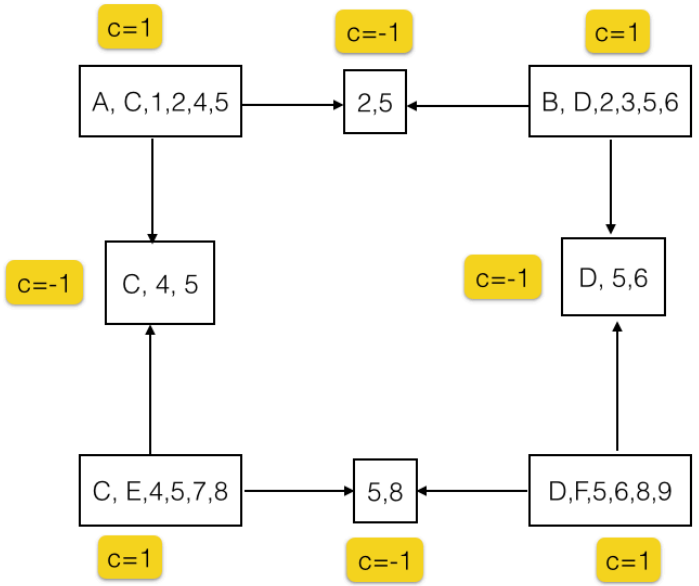
$$\sum_{R \in \mathcal{R}} c_R \mathrm{I}(a \in F_R) = \sum_{R \in \mathcal{R}} c_R \mathrm{I}(i \in V_R) = 1$$

  $\Rightarrow$ ensures that the region graph average energy is exact

- Regions containing a particular variable node, form a connected subgraph
  $\Rightarrow$ Marginal probability is consistent.

Example of not valid region graph

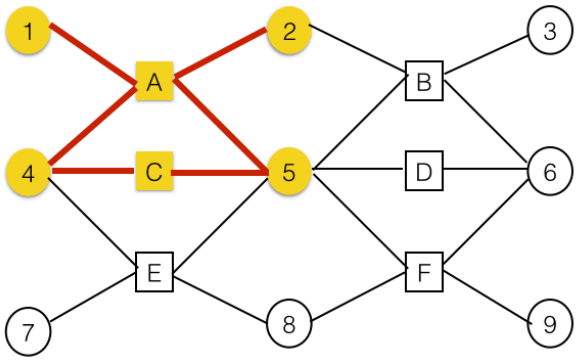


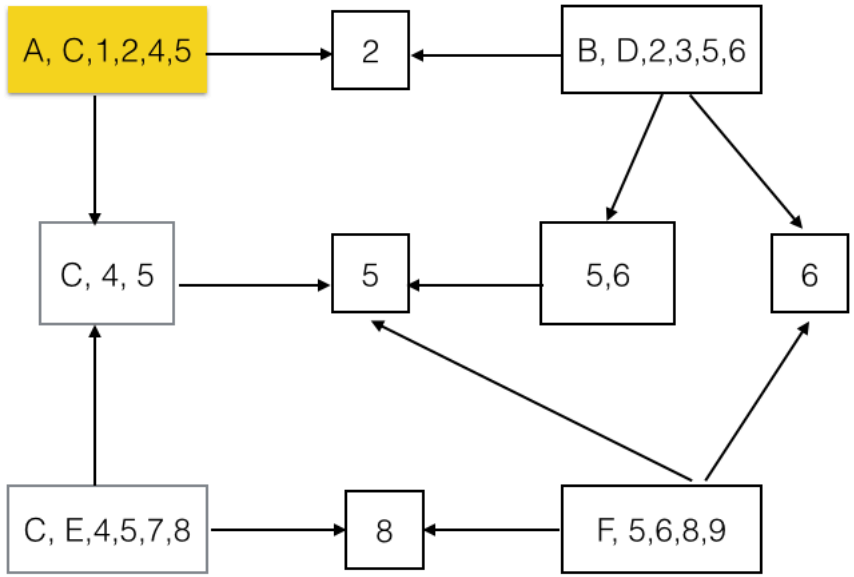


- This is not a valid region graph. Variable 5 is not counted once.

Example of valid region graph

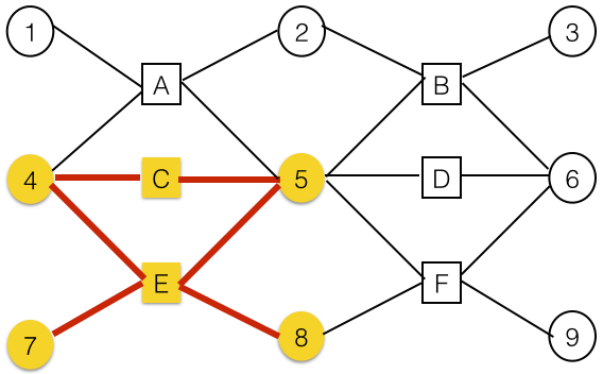
- Bethe region graph for the following factor graph

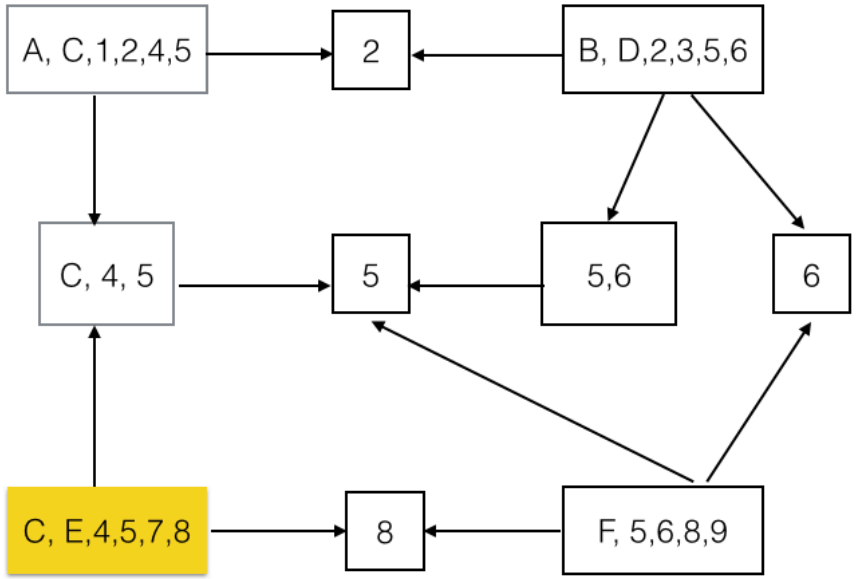




Example of valid region graph

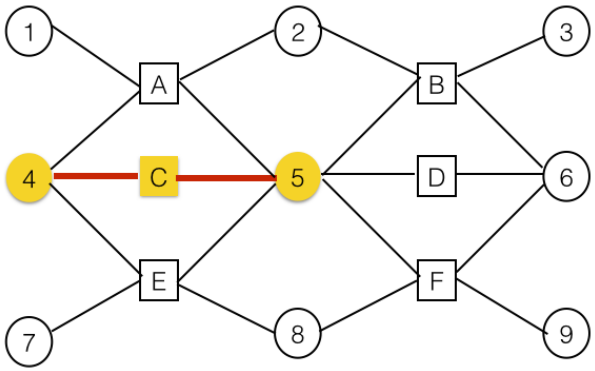
- Bethe region graph for the following factor graph

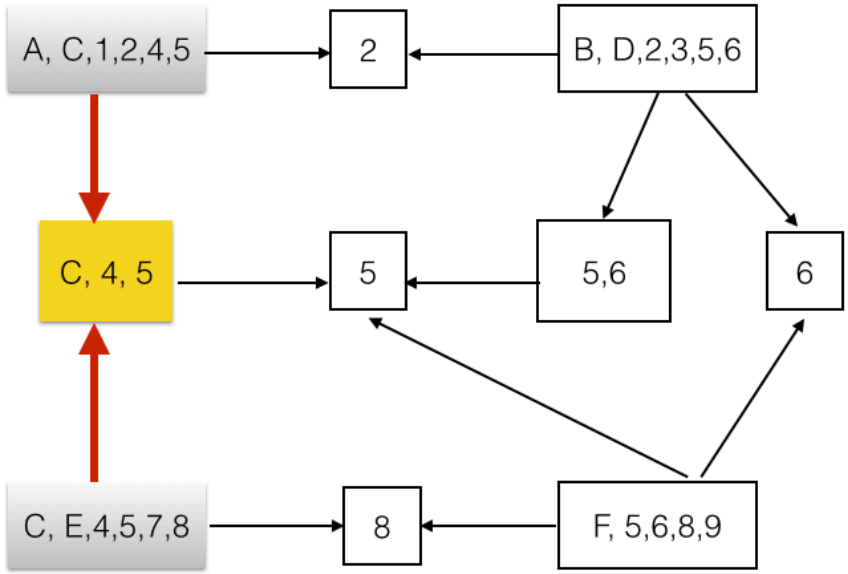




Example of valid region graph

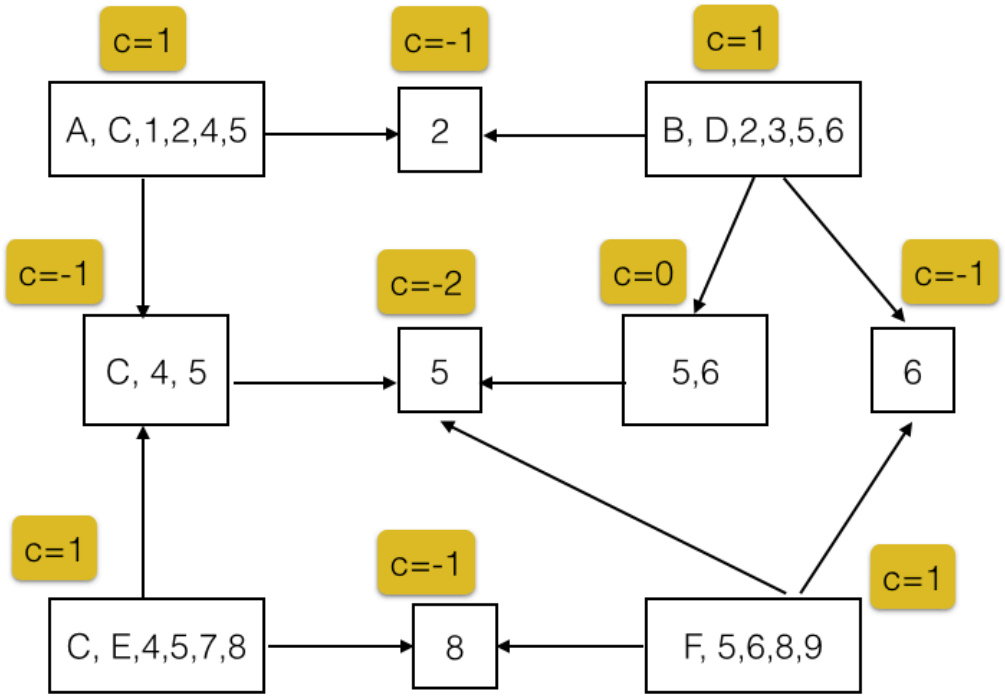
- Bethe region graph for the following factor graph





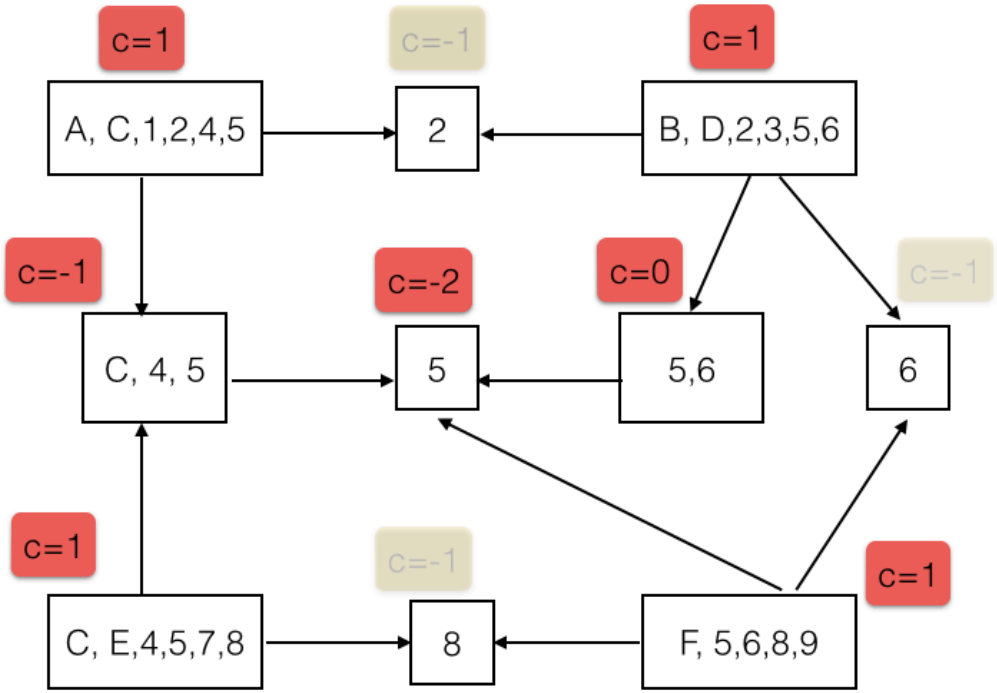
Region graph

$$c_R = 1 - \sum_{R' \in A(R)} c'_R \text{ for all } R \in \mathcal{R}$$



Region graph

- Valid region graph (every node is counted once)



Generalized Belief Propagation

- Theorem: The stationary points of the constrained region-based free energy for a valid region graph, are the fixed points of Generalized belief propagation" for that region.

Stationary point of $F_{\mathcal{R}}(\{b_R\}) = \sum_{R \in \mathcal{R}} c_R F_R(b_R)$

subject to $\sum_{x_R} b_R(x_R) = 1 \quad \text{forall} \quad R \in \mathcal{R}$

$\sum_{x_P \setminus x_C} b_P(x_P) = b_C(x_C) \quad \text{Parent, Child regions} \in \mathcal{R}$

$b_R(x_R) \geq 0$

Generalized Belief Propagation

- Belief in a region is product of:
  - Local information (factors in region)
  - Messages from parent regions
  - Messages into descendant regions from parents who ware not descendant.
- Message update rules obtained by enforcing marginalization constraints.

Generalized Belief Propagation

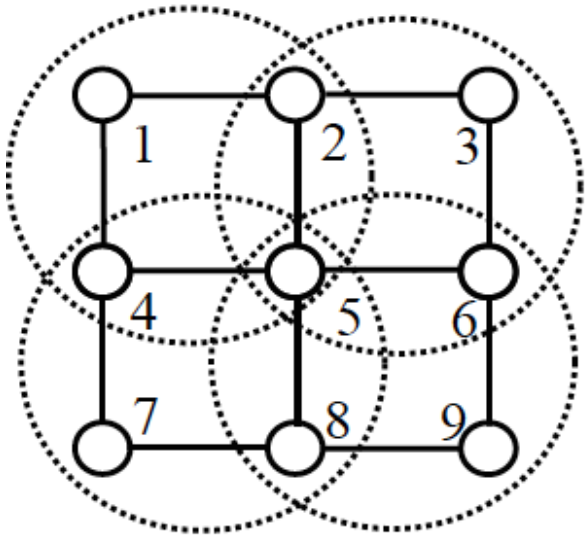
Belief in a region is:

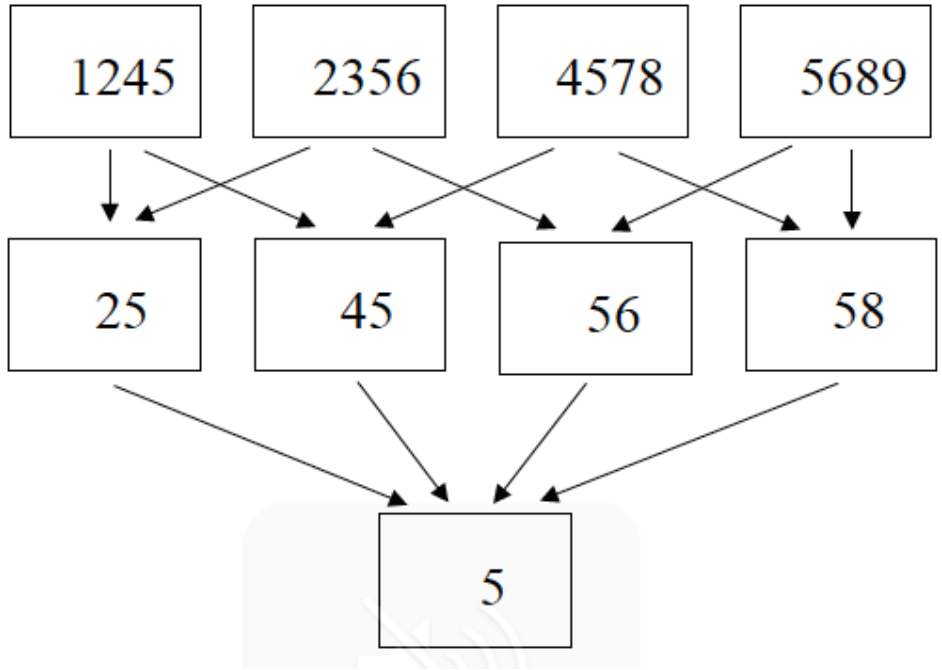
$$b_R(x_R) \propto \prod_{a \in A_R} f_a(x_a) \times \underbrace{\left(\prod_{P \in \mathcal{P}(R)} m_{P \to R}(x_R) \right)}_{\text{Messages from parent regions}}$$

$$\times \underbrace{\left(\prod_{D \in \mathcal{D}(R)} \prod_{P' \in \mathcal{P}(D) \setminus \varepsilon(R)} m_{P' \to D}(x_D) \right)}_{\text{messages into descendant regions from parents who ware not descendant}}$$

Generalized Belief Propagation

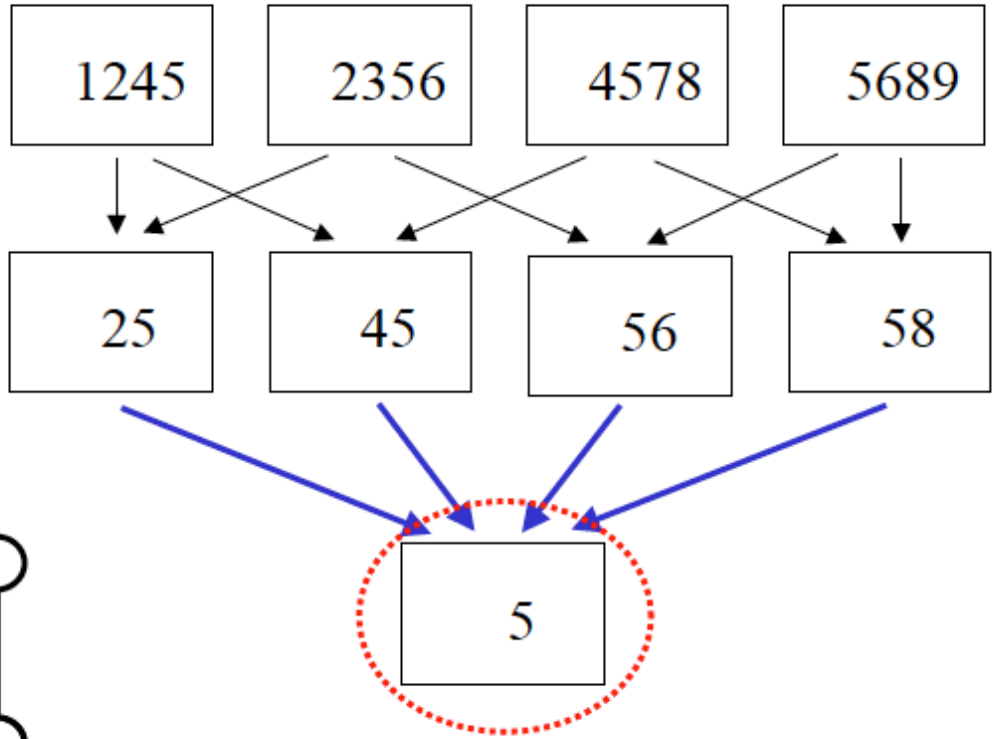
- Bethe region graph for the following graph

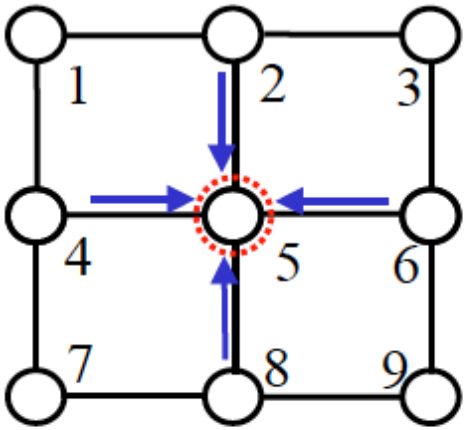




[2]J.S. Yedidia, Construction free energy approximation, 2005

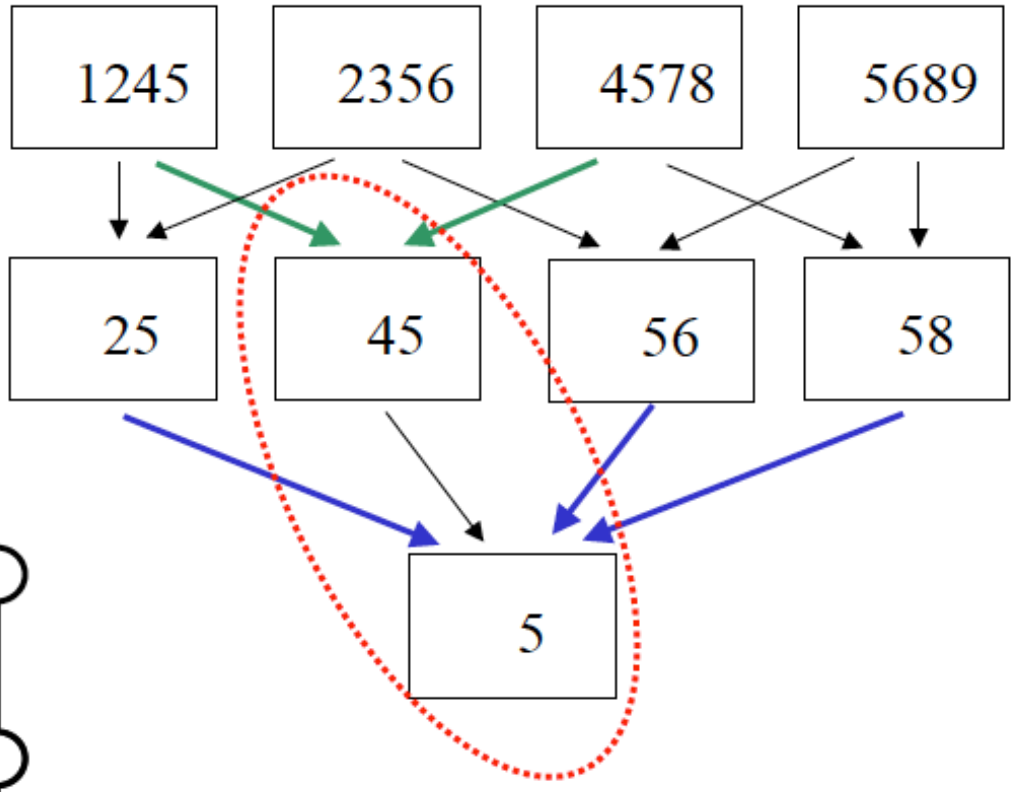
Generalized Belief propagation

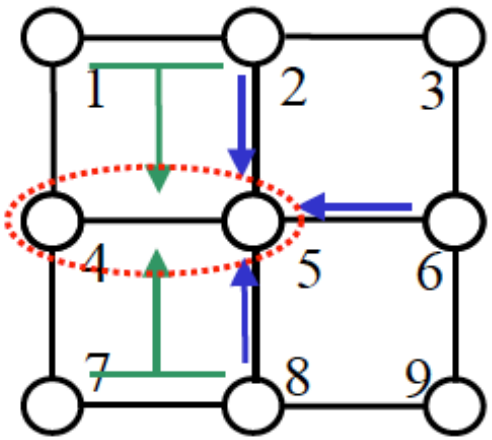




$$b_5 \propto m_{2\to5} m_{4\to5} m_{6\to5} m_{8\to5}$$

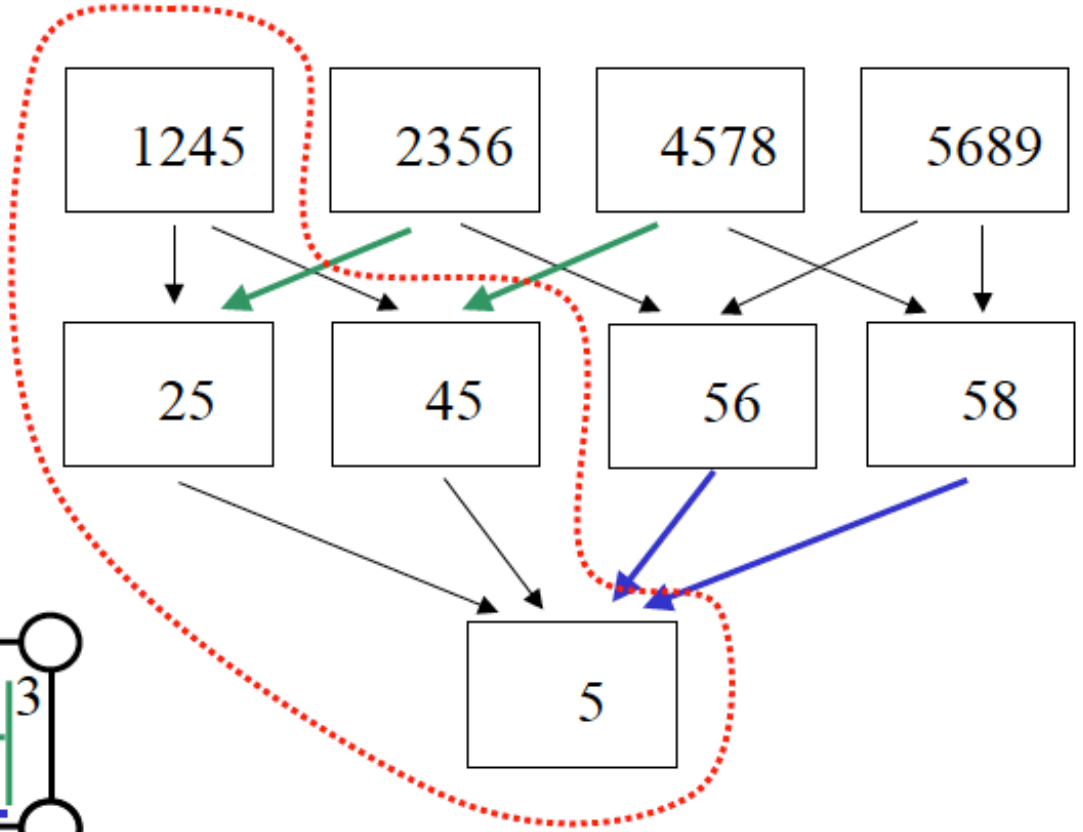
Generalized Belief propagation

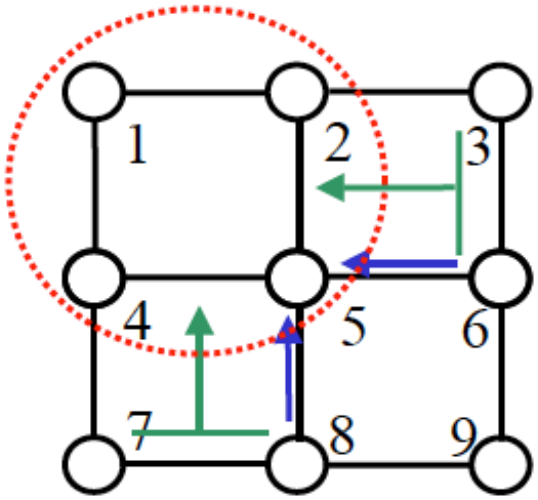




$$b_{45} \propto [f_{45}][m_{12\to45} m_{78\to45} m_{2\to5} m_{6\to5} m_{8\to5}]$$

Generalized Belief propagation

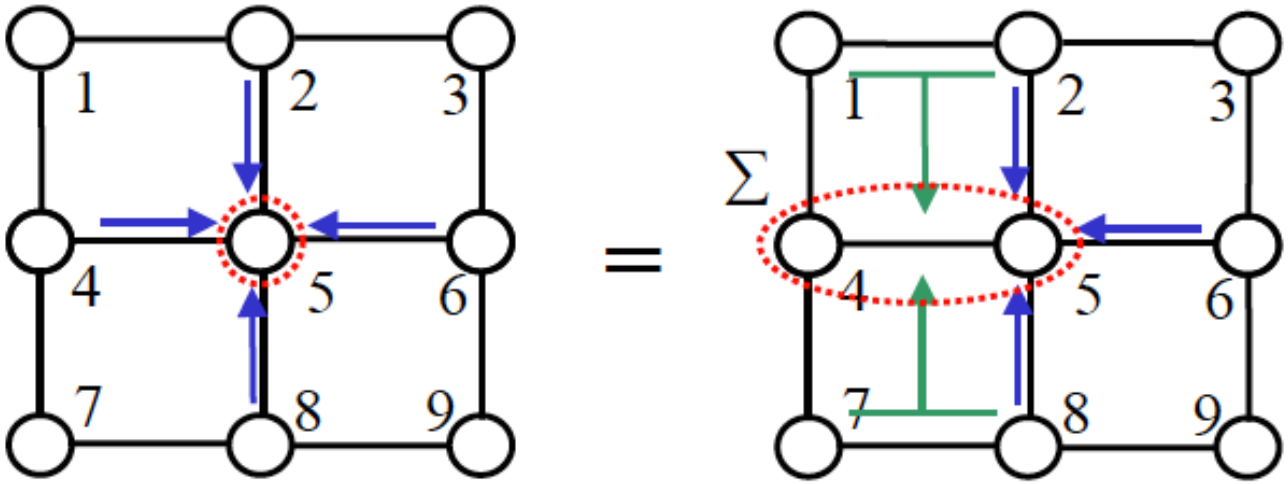




$$b_{1245} \propto [f_{12} f_{14} f_{25} f_{45}][m_{36\to 25} m_{78\to 45} m_{6\to 5} m_{8\to 5}]$$

Generalized Belief propagation

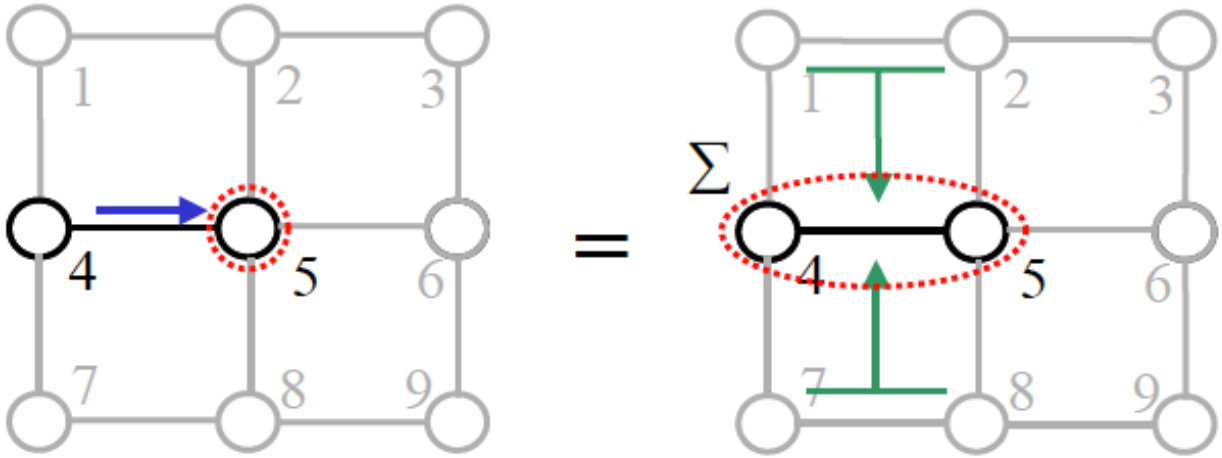
Use marginalization constraints to derive message-update rules



$$b_5(x_5) = \sum_{x_4} b_{45}(x_4, x_5)$$

Generalized Belief propagation

Use marginalization constraints to derive message-update rules



$$m_{4\to 5}(x_5) \propto \sum_{x_4} f_{45}(x_4, x_5) m_{12\to 45}(x_4, x_5) m_{78\to 45}(x_4, x_5)$$

Thanks

Questions?