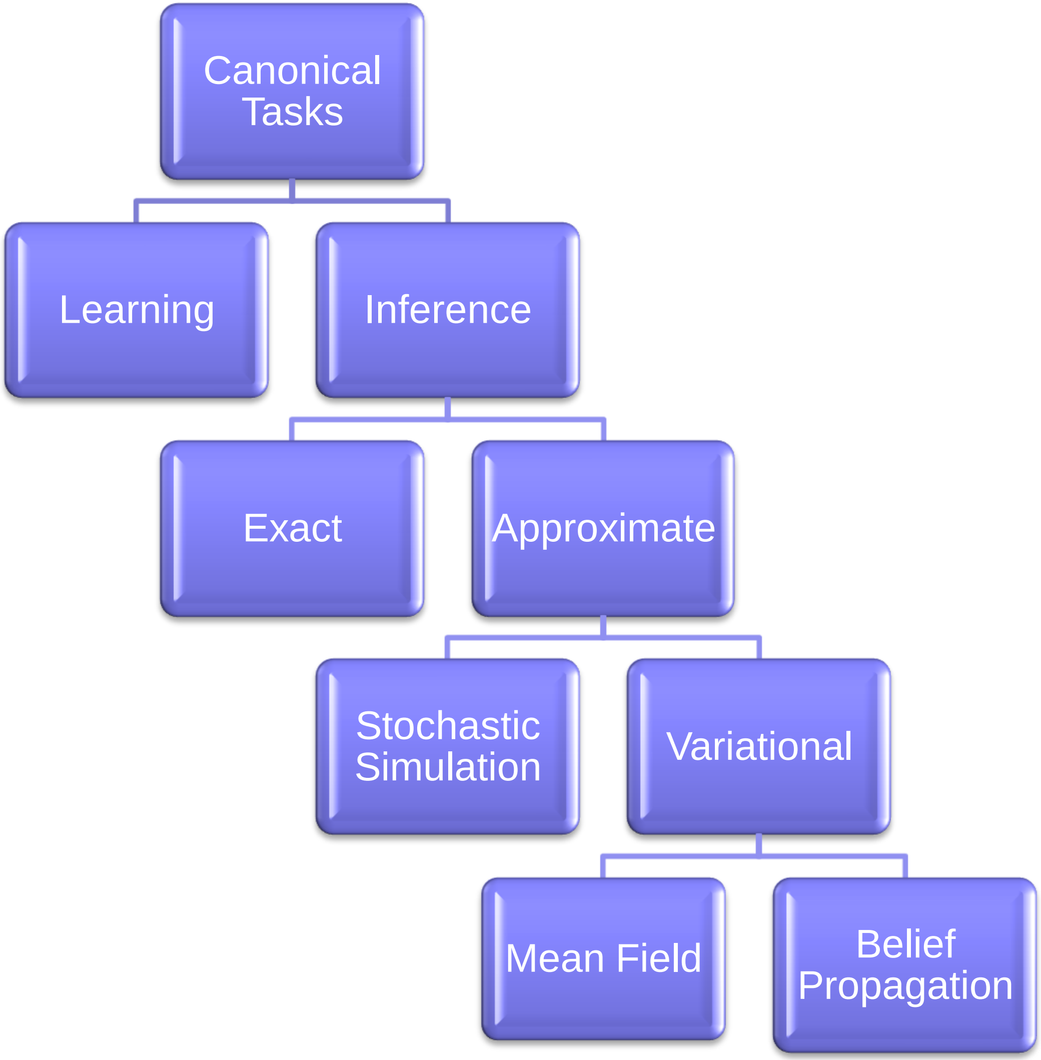
Constructing Free-Energy Approximations and Generalized Belief Propagation Algorithms

Jonathan S. Yedidia William T. Freeman Yair Weiss

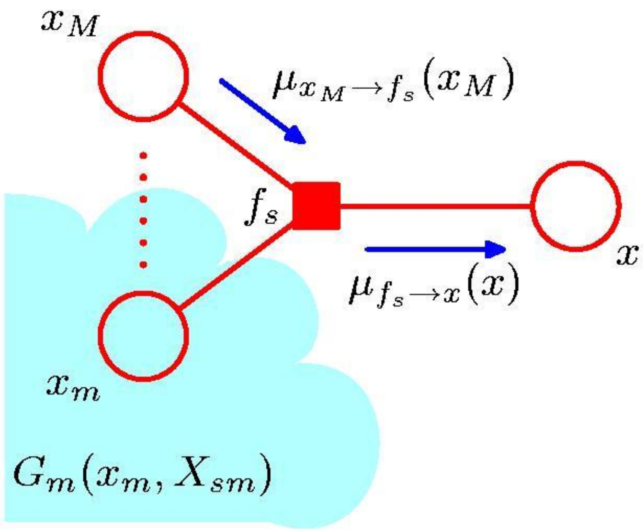
Presented by: Brandon Mayer

March 17, 2010

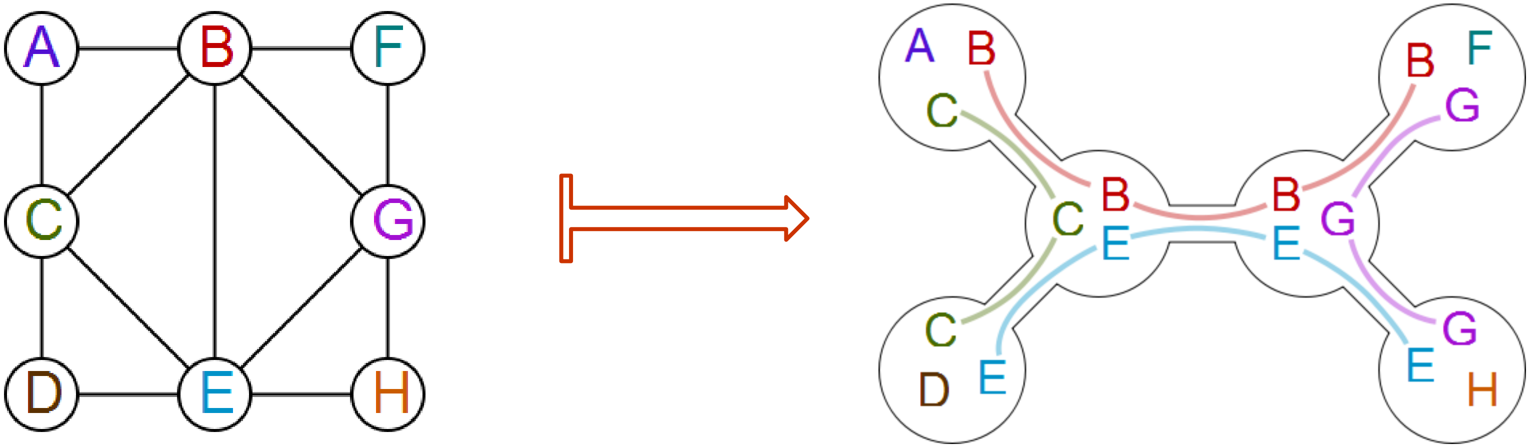
Overview


Canonical Tasks
Learning
Inference
Exact
Approximate
Stochastic Simulation
Variational
Mean Field
Belief Propagation

Standard Belief Propagation



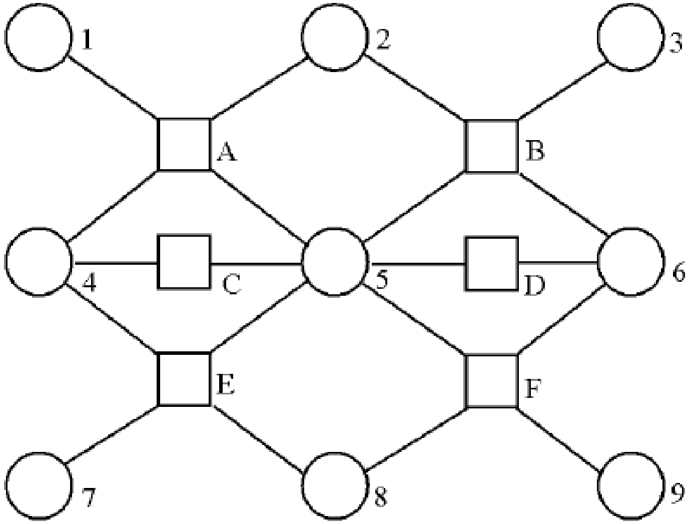
- Sum Product Exact when graph is a tree or can be expressed as a tree.
- Marginalization calculated as a propagation of local products and sums



Standard Belief Propagation

Messages derived for Sum-Product Algorithm are independent of Topology

Can apply Sum-Product to an arbitrary graph (Standard BP)

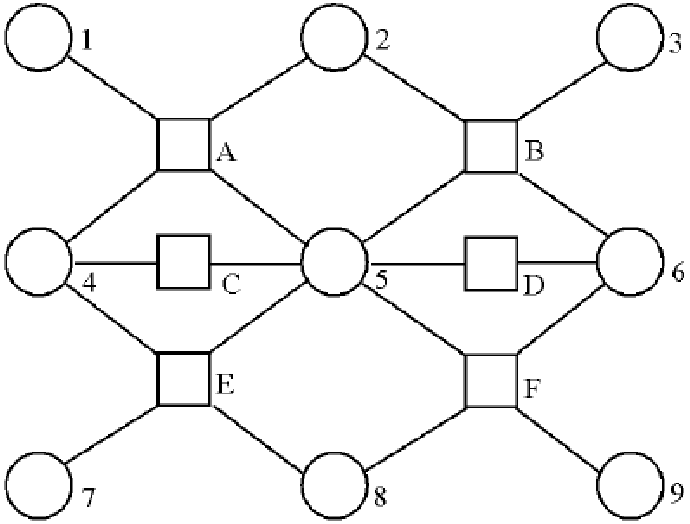


- No Convergence Guarantees
- When BP converges it is usually to very accurate approximations
  - □ Why?
  - □ Is there a way to further generalize BP?

Standard Belief Propagation

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{a} f_a(\mathbf{x}_a)$$

$$b_i(x_i) \approx p_i(x_i)$$



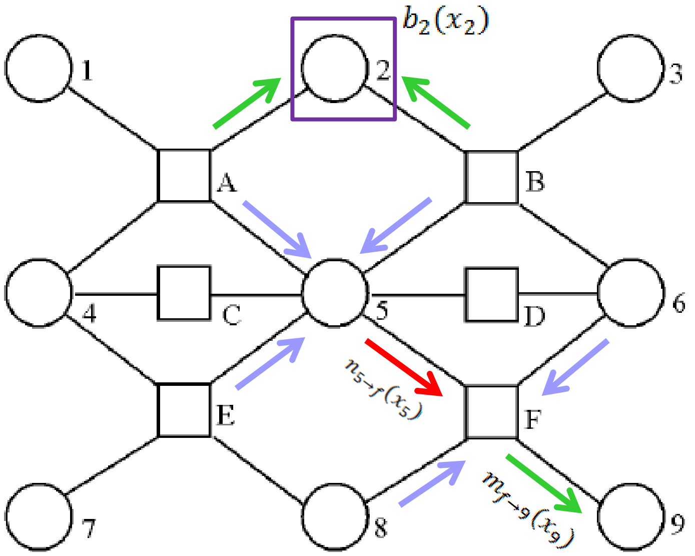
- Consistency (marginalization)

$$b_i(x_i) = \sum_{\mathbf{x}_a \setminus x_i} b_a(\mathbf{x}_a)$$

- Normalization

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

Standard Belief Propagation



Two Types of Messages

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

$$m_{a\to i}(x_i) := \sum_{\mathbf{x}_a\setminus x_i} f_a(\mathbf{x}_a) \prod_{j\in N(a)\setminus i} n_{j\to a}(x_j)$$

Marginal Belief

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

Why is BP effective?

- Free Energy Perspective

Boltzmann's Law

$$p(\mathbf{x}) = \frac{1}{Z(T)} e^{-E(\mathbf{x})/T}$$

Probability of a state of a system in thermodynamic equilibrium

Define:

$$E(\mathbf{x}) = -\sum_{a=1}^{M} \ln f_a(\mathbf{x}_a) \xRightarrow{T\,=\,1} p(\mathbf{x}) = \frac{1}{Z} \prod_a f_a(\mathbf{x}_a)$$

- By this definition of energy we may relate the probability distribution specified by a factor graph to free energies of statistical physics

Why is BP effective?

- Minimize KL divergence between Approximation and Exact Distribution

$$\min_{b(x)} D\big(b(x)||p(x)\big) = \sum_{\{x\}} b(\{x\})\ln\left(\frac{b(\{x\})}{p(\{x\})}\right)$$

Decompose KL into a function of energy by substitution of Boltzmann's law for p({x})

Gibbs Free Energy

$$D\left(b\{x\}||p(\{x\})\right) = \boxed{\boxed{\sum_{\{x\}} b(\{x\})E(\{x\})} + \boxed{\sum_{\{x\}} b(\{x\})\ln b(\{x\})}} + \ln Z$$

Average Energy

Negative Entropy

$$F(b) = U(b) - H(b)$$

Helmholtz Free Energy

$$\min_{b(x)} D(b(x)||p(x)) \Rightarrow F(b) = \boxed{F_H = -\ln Z}$$

Why is BP effective?

- Can minimize F(b) w.r.t to b(x) to recover p(x)
  - Usually intractable so we minimize b(x) with a given form

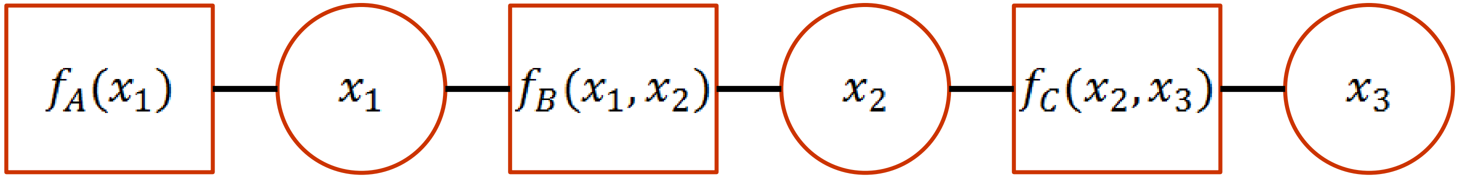
Mean Field Approximation

$$\boxed{b_{MF}(\mathbf{x}) = \prod_{i=1}^{N} b_i(x_i)}$$

More interesting to specify an approximation which depends on pairs of nodes

Why is BP effective?

Example: consider a Markov Chain



$$p(\boldsymbol{x}) = \frac{p(x_1)p(x_1,x_2)p(x_2,x_3)}{p(x_1)p(x_2)} = p(x_1)p(x_2|x_1)p(x_3|x_2)$$

In General an a singly connected factor graph’s joint probability distribution can be represented as:

$$\boxed{p(\mathbf{x}) = \frac{\prod_{a=1}^{M} p_a(\mathbf{x}_a)}{\prod_{i=1}^{N} \left(p_i(x_i)\right)^{d_i - 1}}}$$

$$d_i \coloneqq number\ of\ factor\ nodes\ connected\ to\ i^{th}\ variable\ node$$

Why is BP effective?

We may specify a family of belief approximations which would be exact if the factor graph was without cycles

$$b(\boldsymbol{x}) = \frac{\prod_{a=1}^{M} p_a(x_a)}{\prod_{i=1}^{N} \left(p_i(x_i)\right)^{d_i - 1}}$$

Then substitute this definition of the joint belief into the variational Entropy to define:

$$H_{Bethe} = -\sum_{a=1}^{M} \sum_{\boldsymbol{x}_a} b_a(\boldsymbol{x}_a) \ln b_a(\boldsymbol{x}_a) + \sum_{i=1}^{N} (d_i - 1) \sum_{x_i} b_i(x_i) \ln b_i(x_i)$$

We may define the variational Average Energy as:

$$U_{Bethe} = -\sum_{a=1}^{M} \sum_{\boldsymbol{x}_a} b_a(\boldsymbol{x}_a) \ln f_a(\boldsymbol{x}_a)$$

Which is exact when b(x) = p(x)

Why is BP effective?

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

To determine the value of b(x) which minimized D(b||p) we construct a Lagrangian, L to enforce consistency and normalization conditions

$$L\left(b(\boldsymbol{x}), \lambda_{a,i}, \lambda_i\right) = F_{Bethe} + \sum_i \lambda_i h_i + \sum_{a,i} \lambda_{a,i} g_{a,i}$$

Minimizing the Bethe Energies under this approximation we recover the standard Belief Propagation fixed-point belief equations

$$\hat{b}_a(\mathbf{x}_a) \propto f_a(\mathbf{x}_a) \prod_{i \in N(a)} \prod_{c \in N(i) \backslash a} m_{c \to i}(x_i)$$

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

Why is BP effective?

- Belief Propagation fixed points are equivalent to local minima of the Bethe free energy approximation.
- Standard BP attempts to minimize the Bethe approximation of the Gibbs Free Energy
  - If the graph does not have cycles the Bethe approximation becomes exact.
  - Can be a very good approximation in the presence of cycles.

Questions

1. Can we develop better approximations to the Gibbs free energy F(b)?
2. Can we develop message passing algorithms that correspond to stationary points of the improved approximations?

Generalizing BP

Entropy Approximation

$$H(x,y) = H(x) + H(y) - I(x,y)$$

$$I(x,y) = \sum_x \sum_y p(x,y) \ln\left(\frac{p(x,y)}{p(x)p(y)}\right) = D(p(x,y)||p(x)p(y))$$

Bethe approximation includes higher order terms in variational entropy and is exact when graph is acyclic but an approximation on a graph with cycles.

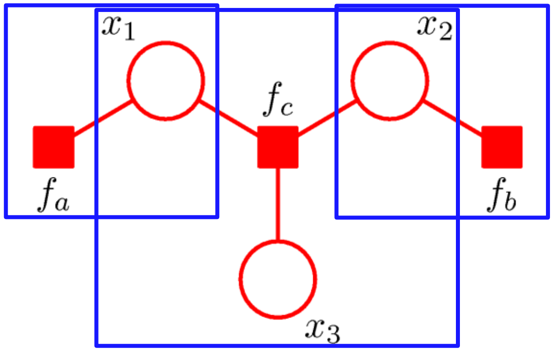
We may construct better approximations to the variational entropy by considering higher order contributions to the mutual information term.

Generalizing BP

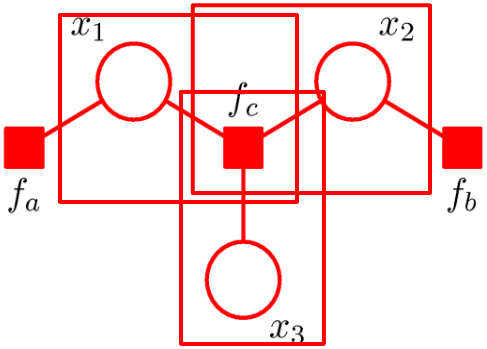
Define Belief approximations over larger sets of nodes (Regions) in stead of fully factored (MF) or pairwise (Bethe) nodes.

Region R is a set of variable and factor nodes such that if a factor node **a** is included in the region all variable nodes connected to **a** must be included in the region

Valid Regions



Invalid Regions



$$E_R(\boldsymbol{x}_R) = -\sum_{a \in A_R} \ln f_a(\boldsymbol{x}_a) := \mathit{Region\ Energy}$$

$$U_R(b_r) = \sum_{\boldsymbol{x}_R} b_R(\boldsymbol{x}_r) E_R(\boldsymbol{x}_R) := \mathit{Region\ Average\ Energy}$$

$$H_R(b_R) = -\sum_{\boldsymbol{x}_R} b_R(\boldsymbol{x}_R) \ln b_R(\boldsymbol{x}_R) := \mathit{Region\ Entropy}$$

Advantage: Add higher terms to variational entropy approximation. Express global energy approximation as a sum over local regions

Disadvantage: Computation Costs

Generalizing BP

Definition: Valid Region Based Free Energy Approximation

For a region based free energy approximation to be considered valid, we must ensure every factor and variable node will be counted exactly one time in the energy approximation.

Definition: Maxent Normal

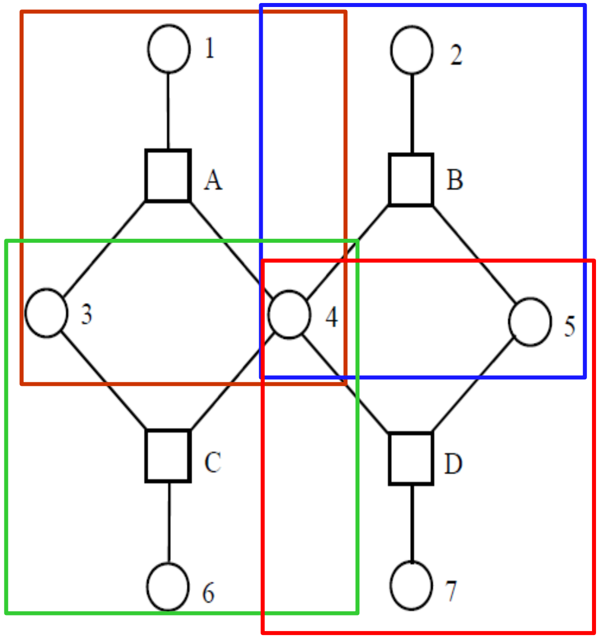
An approximation is considered maxent normal when the variational entropy is maximum if the beliefs are uniform.

The Bethe approximation (standard BP) is always a valid region based energy approximation and is always maxent normal

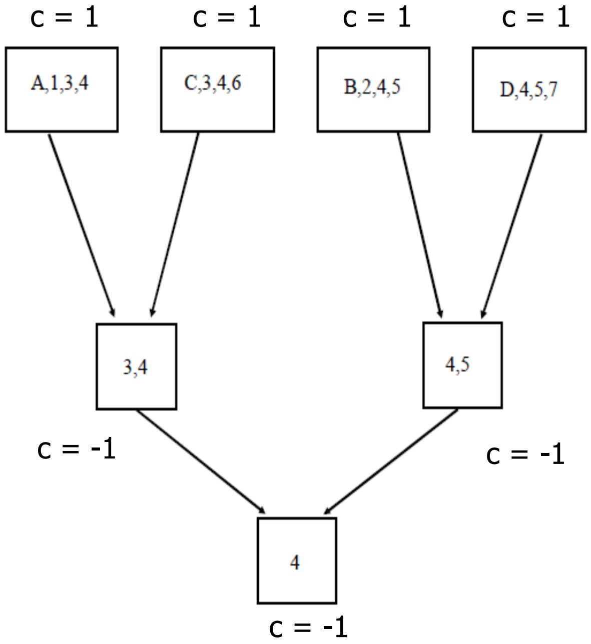
The methods described hereafter are always valid energy approximations but are not guaranteed to be maxent normal

Generalizing BP: Region Graph Method

- Each Vertex is a set of valid regions
  - Set of factor and variable nodes
- A directed edge may exist between two vertices if the child vertex is a subset of the parent
- Define counting numbers to ensure the region graph will yield a valid energy approximation.



$$c_v = 1 - \sum_{u \in A(v)} c_u$$


c = 1 c = 1 c = 1 c = 1

A,1,3,4 C,3,4,6 B,2,4,5 D,4,5,7

3,4 4,5

c = -1 c = -1

4

c = -1

$$\boxed{F_{RG} = \sum_R c_R F_R}$$

Note:

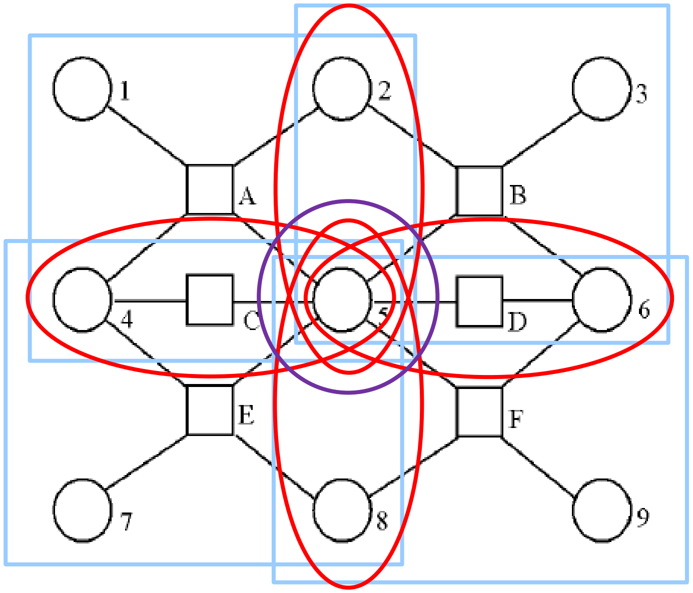
$$\sum_{v \in V(i)} c_v = 1$$

Cluster Variation

- Cluster Variational Method (Kikuchi)
  - □ Method for selecting a valid set of regions $\mathcal{R}$ and counting numbers c_r
  - □ Free energy is approximated as a sum of the local free energies of basic clusters minus the intersection of over-counted nodes, minus the intersection of these intersections and so on…
  - □ Bethe approximation is a special case of choosing pair-wise clusters iff no factor nodes shares more than one variable node with another factor node.

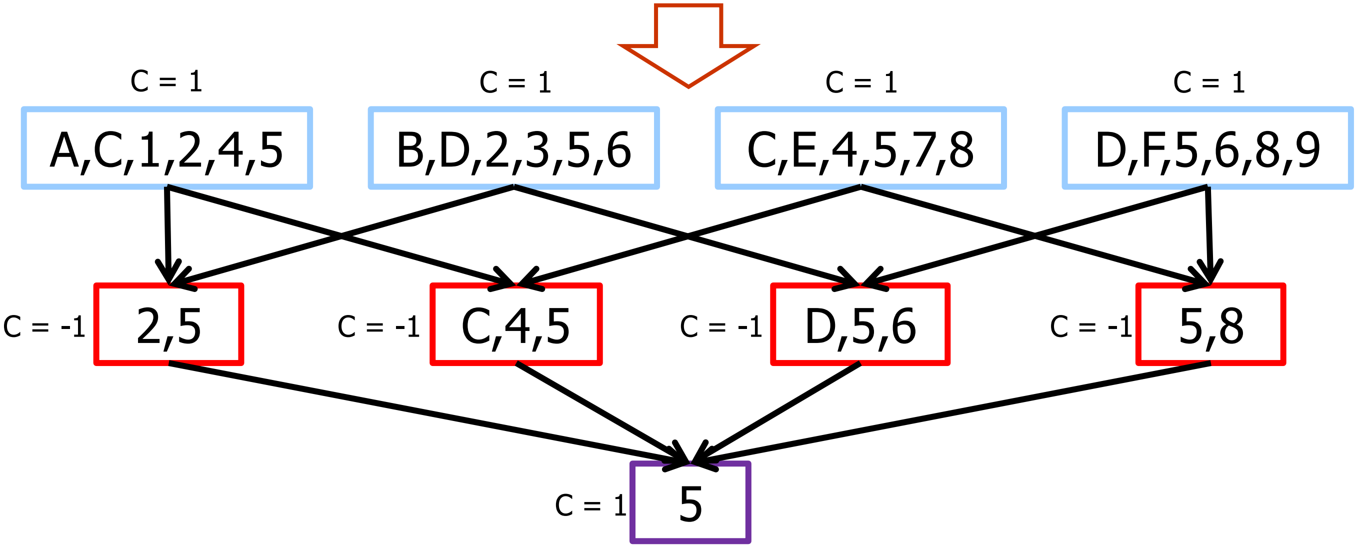
1. Choose initial set of clusters $\mathcal{R} \in R_0$ such that every variable and factor node is included in at least one region $\mathcal{R} \in R_0$
2. For every subsequent level $\mathcal{R}_1, \mathcal{R}_2, \dots, \mathcal{R}_K$ construct regions which consist of intersections between its previous levels discarding intersections that appeared in previous level or are sub-regions of regions in the same level
3. Connect all regions in a $R \in \mathcal{R}_j$ to super-regions of previous levels $\mathcal{R} \in \mathcal{R}_{l<j}$ unless the region $\mathcal{R} \in \mathcal{R}_{l<j}$ is an ancestor of a region already connected to $R \in \mathcal{R}_j$

Cluster Variation



$$c_R = 1 - \sum_{S \in \mathcal{S}(R)} c_S$$

$S(R) :=$ *Super – Region of R*



C = 1 A,C,1,2,4,5

C = 1 B,D,2,3,5,6

C = 1 C,E,4,5,7,8

C = 1 D,F,5,6,8,9

C = -1 2,5

C = -1 C,4,5

C = -1 D,5,6

C = -1 5,8

C = 1 5

Junction Graphs

Bipartite Directed Graph

$$G_{JG} = (V_L, V_S, E, L)$$

$$V_L \coloneqq Large\ Vertices$$

$$V_S \coloneqq Small\ Vertices$$

$$L \coloneqq Vertex\ Labels$$

$$N(v_s(i)) \coloneqq Neighborhood\ of\ the\ i^{th} small\ vertex$$

Construction Rules

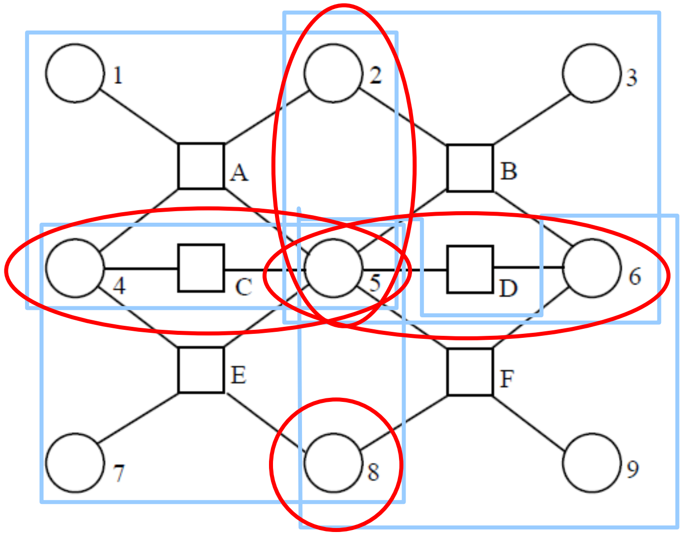
$$if\ N(v_s) = \{v_{l1}, v_{l2}, \dots, v_{lk}\}$$
$$l(v_s) \subseteq l(v_{l_1}) \cap l(v_{l_2}) \cap \dots \cap l(v_{l_k})$$

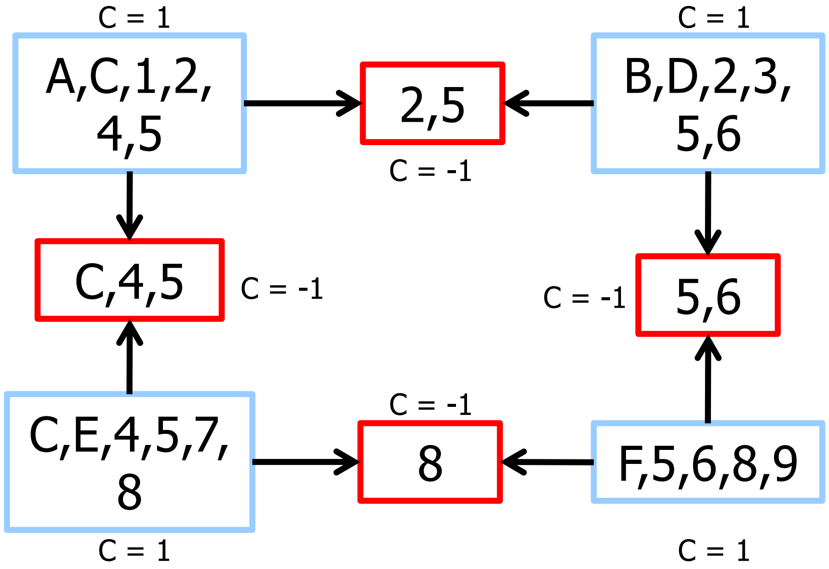
For each i, the subgraph of G consisting only of the vertices which contain i in their labels must be a connected tree

$c_r = 1\ \forall\ V_L$

$c_r = 1 - d_r\ \forall\ V_S$

$d_r \coloneqq number\ of\ neighboring\ large\ regions$




C = 1
A,C,1,2,4,5

2,5
C = -1

C = 1
B,D,2,3,5,6

C,4,5
C = -1

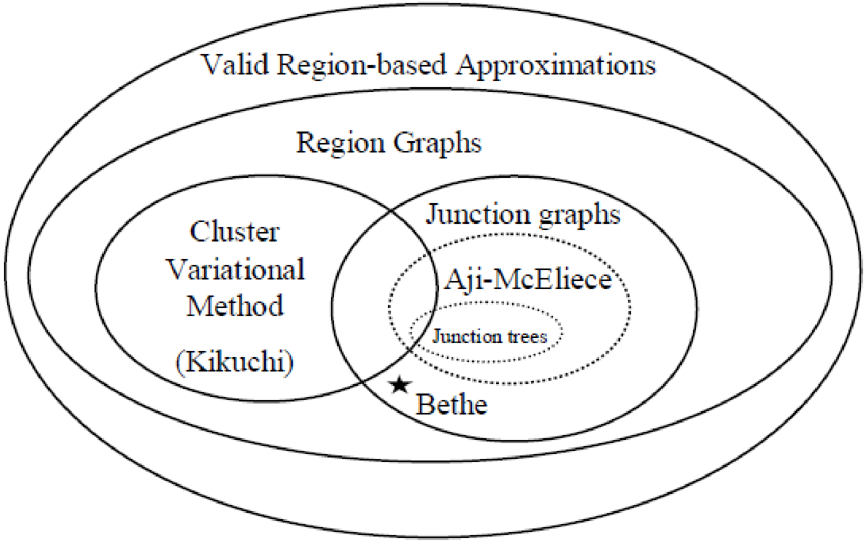
C = -1
5,6

C,E,4,5,7,8
C = 1

C = -1
8

F,5,6,8,9
C = 1

Relationship between Region Graphs



- There exists Region graphs that cannot be produced by Cluster Variational or Junction Graphs methods.

- There also exist valid region based approximations that cannot be produced by the region graph method

- However the region graph method allows for a convenient generalization for GBP message passing so we utilize this framework

Generalizing BP: Message Passing

Given a region graph there are several GBP message passing algorithms

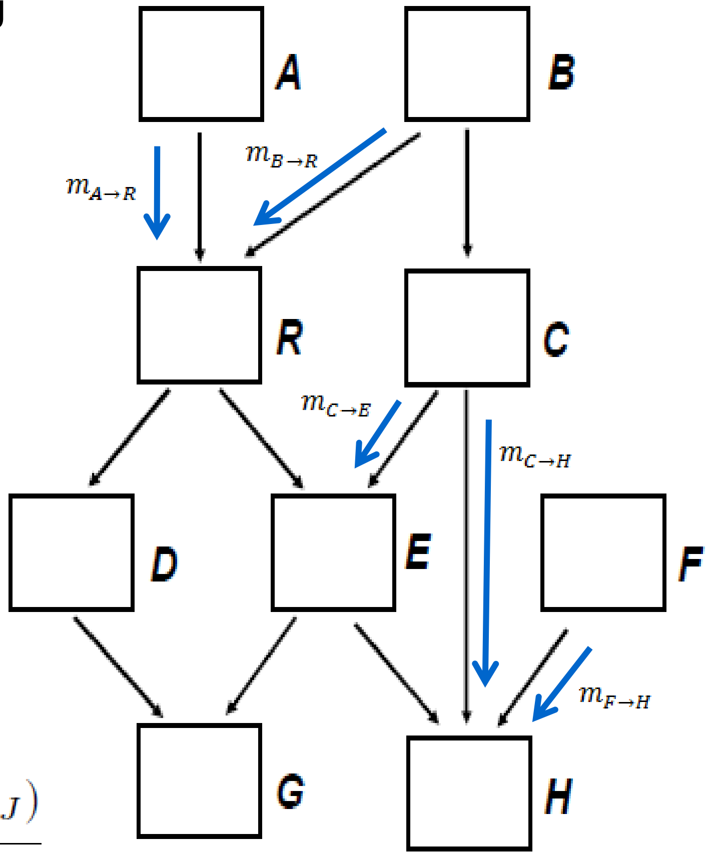
Ex: Parent-to-Child Algorithm

Main Advantage: No need to consider counting numbers in messages

Rules:

1. Belief at region R is the product of messages from R's parents and messages to R's descendants from their parents excluding R and its descendants.

$$b_R \propto m_{A\to R}\, m_{B\to R}\, m_{C\to E}\, m_{C\to H}\, m_{F\to H} \prod_{a\in A_R} f_a(\mathbf{x}_a)$$



Enforcing Consistency:

$$m_{P\to R}(x_R) := \frac{\sum_{x_{P\setminus R}} \prod_{a\in F_{P\setminus R}} f_a(x_a) \prod_{(I,J)\in N(P,R)} m_{I\to J}(x_J)}{\prod_{(I,J)\in D(P,R)} m_{I\to J}(x_J)}$$

Example on next slide

$$E(R) = R \cup D(R)$$

$$N(P,R) := \{I,J \mid J \in E(P), J \notin E(R), I \notin E(P)\}$$

$$D(P,R) := \{I,J \mid J \in E(R), I \in E(P), I \notin E(R)\}$$

Generalizing BP

Example of enforcing consistency to derive message updates

$$\boxed{b_{1,2}(x_1,x_2) \propto f_1 f_2 m_{35\to12} m_{46\to12} m_{3\to1} m_{4\to1}}$$

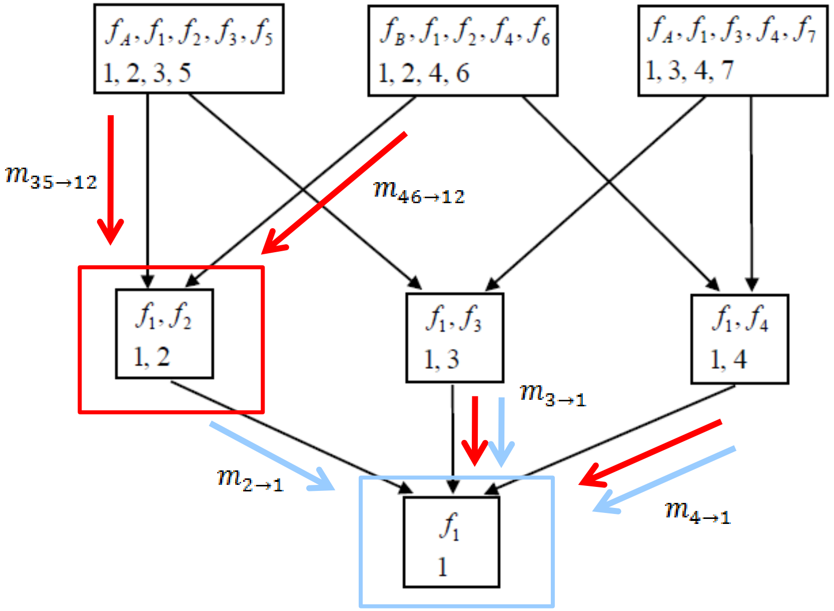
$$\boxed{b_1(x_1) \propto f_1 m_{2\to1} m_{3\to1} m_{4\to1}}$$

Enforce Consistency

$$b_1(x_1) = \sum_{\forall x_2} b_{1,2}(x_1,x_2)$$

$$f_1 m_{2\to1} m_{3\to1} m_{4\to1} = \sum_{\forall x_2} f_1 f_2 m_{35\to12} m_{46\to12} m_{3\to1} m_{4\to1}$$

$$\boxed{m_{2\to1} = \sum_{\forall x_2} f_2 m_{35\to12} m_{46\to12}}$$



Generalizing BP

Example of enforcing consistency to derive message updates

$$\boxed{b_{1,2}(x_1,x_2) \propto f_1 f_2 m_{35\to12} m_{46\to12} m_{3\to1} m_{4\to1}}$$

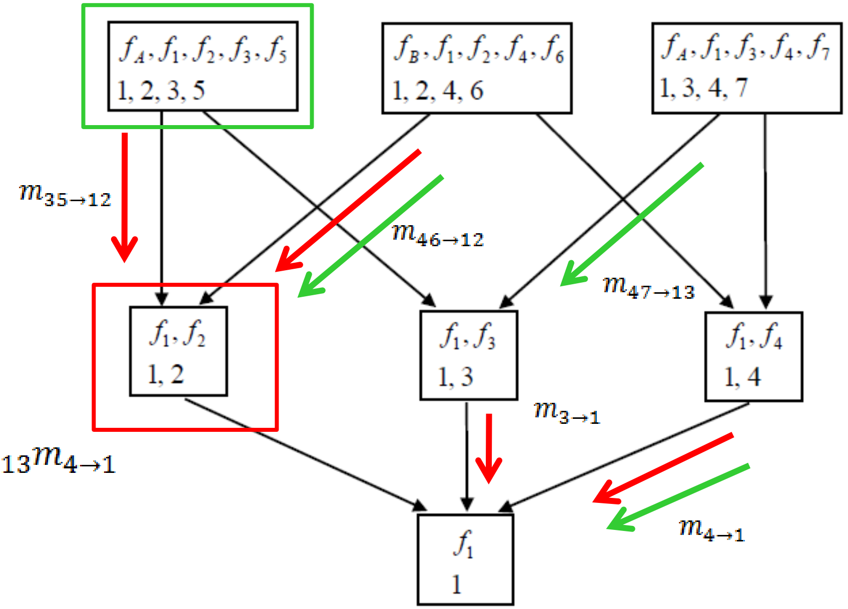
$$\boxed{b_{1235}(x_1,x_2,x_3,x_5) \propto f_a f_1 f_2 f_3 f_5 m_{46\to12} m_{47\to13} m_{4\to1}}$$

Enforce Consistency

$$b_{1,2}(x_1,x_2) = \sum_{x_3,x_4} b_{1,2,3,5}(x_1,x_2,x_3,x_5)$$

$$f_1 f_2 m_{35\to12} m_{46\to12} m_{3\to1} m_{4\to1} = \sum_{x_3,x_4} f_a f_1 f_2 f_3 f_5 m_{46\to12} m_{47\to13} m_{4\to1}$$

$$\boxed{m_{35\to12} = \frac{\sum_{x_3,x_4} f_a f_3 f_5 m_{47\to13} m_{4\to1}}{m_{3\to1}}}$$



There exist other GBP algorithms that modify the message passing dynamics but which yield the same solutions. (Child-to-Parent, Two-Way Algorithm)

Summary

- Fixed Points of Standard BP are local minima of the Bethe Free Energy approximation
  - No convergence guarantees
  - May have multiple fixed points
- Use region graph method to construct more general energy approximations
  - Cluster Variation Method
  - Junction Graph Method
  - Can guarantee valid energy approximation
  - Cannot Guarantee maxent-normality
  - No Convergence/Performance guarantees
- General BP message Passing Algorithms
  - Different types of GBP message passing algorithms with different advantages
  - The fixed points of the GBP message passing algorithms correspond to local minima of the region based energy approximation given a choice of region graph.

References

- Constructing Free-Energy Approximations and Generalized Belief Propagation Algorithms
  - Jonathan S. Yedidia, William T. Freedman, Yair Weiss
- Understanding Belief Propagation and its Generalizations
  - Jonathan S. Yedidia, William T. Freedman, Yair Weiss
- Pattern Recognition and Machine Learning
  - Christopher M. Bishop
- Probabilistic Graphical Models
  - Daphne Koller, Nir Friedman