Why BP Works

STAT 232B

Free Energies

— Helmholz & Gibbs Free Energies (1)

- Distance between Probabilistic Models - K-L divergence

$$KL(b(\{x\}) \,\|\, p(\{x\}) = \sum_{\{x\}} b(\{x\}) \ln \frac{b(\{x\})}{p(\{x\})}$$

Here, p({x}) is the exact joint prob. b({x}) is the approximation, called "belief".

- Boltzmann's law for computing joint prob.

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

Free Energies

— Helmholz & Gibbs Free Energies (2)

$$KL(b(\{x\}) \,||\, p(\{x\})) = \sum_{\{x\}} b(\{x\}) \ln \frac{b(\{x\})}{p(\{x\})}$$
$$= \sum_{\{x\}} b(\{x\}) \ln b(\{x\}) + \sum_{\{x\}} b(\{x\}) E(\{x\}) / T + \ln Z$$

Since KL(.||.) >= 0,

$$\sum_{\{x\}} b(\{x\}) \ln b(\{x\}) + \sum_{\{x\}} b(\{x\}) E(\{x\}) / T + \ln Z \geq 0$$

Define $F = -T \ln Z$

$$T \sum_{\{x\}} b(\{x\}) \ln b(\{x\}) + \sum_{\{x\}} b(\{x\}) E(\{x\}) \geq F$$

F is called the "Helmholz free energy", which is the lower bound of the above inequality.

Free Energies

— Helmholz & Gibbs Free Energies (3)

Let's define,

$$G(b(\{x\})) = \sum_{\{x\}} b(\{x\}) E(\{x\}) + T \sum_{\{x\}} b(\{x\}) \ln b(\{x\}) = U(b\{x\}) - TS(b\{x\}) \quad (1)$$

where G(b({x})) is called "Approximate Gibbs free energy", U is called "average energy", and S is called "entropy".

The "Exact Gibbs free energy" is defined as

$$G_{exact}(p(\{x\})) = \sum_{\{x\}} p(\{x\}) E(\{x\}) + T \sum_{\{x\}} p(\{x\}) \ln p(\{x\})$$
$$= U(p\{x\}) - TS(p\{x\}) = F$$

The Exact Gibbs free energy is equal to the Helmholtz free energy (at equilibrium).

Free Energies

— Mean-field free energy – a variational approach (1)

Let's introduce an arbitrary "trial" energy function E_α^0, a trial prob. is constructed as:

$$p_\alpha^0 = \frac{\exp(-E_\alpha^0 / T)}{\sum_\alpha \exp(-E_\alpha^0 / T)}$$

Since $$Z = \sum_\alpha \exp(E_\alpha / T)$$

We have

$$Z = \frac{\sum_\alpha \exp(-(E_\alpha - E_\alpha^0)/T) \cdot \exp(-E_\alpha^0 / T)}{\sum_\alpha \exp(-E_\alpha^0 / T)} \cdot \sum_\alpha \exp(-E_\alpha^0 / T)$$

$$= (\sum_\alpha \exp(-(E_\alpha - E_\alpha^0)/T) \cdot p_\alpha^0) \cdot \sum_\alpha \exp(-E_\alpha^0 / T)$$

$$=< \exp(-(E_\alpha - E_\alpha^0)/T) >_0 \cdot \sum_\alpha \exp(-E_\alpha^0 / T)$$

where <.> is the expectation.

Free Energies

— Mean-field free energy – a variational approach (2)

By the property of the convexity of the exponential function :

$$< \exp(-x) > \geq \exp(- < x >)$$

We have $$Z \geq \exp(- < (E_\alpha - E_\alpha^0)/T >_0) \cdot \sum_\alpha \exp(-E_\alpha^0 / T)$$

Then $$F \leq -T \ln \sum_\alpha \exp(-E_\alpha^0 / T) + < (E_\alpha - E_\alpha^0) >_0 \equiv F_{\text{var}}$$

After a few more steps manipulate, we have

$$F_{\text{var}} =< E >_0 - TS_0 \geq F$$

where $$S_0 = -\sum_\alpha p_\alpha^0 \ln p_\alpha^0$$

This suggests us a useful variational arguments: look for the trial prob. func. p_α^0 which gives us the lowest variational free energy. The closer the trial prob. to the exact joint prob., the better the variational approximation.

Free Energies

— Mean-field free energy (3)

Mean-field theory assumes a trial probability func. bearing the factorized form

$$p^0(\{x\}) = \prod_i b_i(x_i)$$

where

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

The energy of a configuration of a pairwise MRF is

$$E(\{x\}) = -\sum_{(i,j)} \ln \psi_{ij}(x_i, x_j) - \sum_i \ln \phi_i(x_i)$$

Plugging this energy into (1), we obtain mean-field Gibbs free energy

$$G_{MF} = U_{MF} - TS_{MF}$$

Free Energies

— Mean-field free energy (4)

where
$$\begin{aligned} U_{MF} &= \sum_{\{x\}} b(\{x\}) E(\{x\}) \\ &= -\sum_{(i,j)} \sum_{x_i, x_j} b_i(x_i) b_j(x_j) \ln \psi_{ij}(x_i, x_j) - \sum_i \sum_{x_i} b_i(x_i) \ln \phi_i(x_i) \end{aligned}$$

and
$$S_{MF} = -\sum_{\{x\}} b(\{x\}) \ln b(\{x\}) = -\sum_i \sum_{x_i} b_i(x_i) \ln b_i(x_i)$$

Note the exact Gibbs free energy is a func. of full joint prob (Helmholz free energy lowest bound of KL divergence). But the mean-field Gibbs free energy is only a func. of the one-node beliefs. To obtain the best approximation of p({x}), we need to search for b({x}) which minimize G_{MF} .

Free Energies

— The Bethe free energy (1)

For tree-like topology MRF, the exact joint prob. can be factorized into a form that only depends on one-node and two-node marginal prob.

$$b(\{x\}) = \prod_{i,j} b_{ij}(x_i, x_j) \prod_i [b_i(x_i)]^{1-q_i}$$

where q_i is the number of nodes that are connected to node i.

We define $E_{ij}(x_i, x_j) = -\ln \psi_{ij}(x_i, x_j) - \ln \phi_i(x_i) - \ln \phi_j(x_j)$

and $E_i(x_i) = -\ln \phi_i(x_i)$

We obtain

$$S_B = -\sum_{(i,j)x_i,x_j} b_{ij}(x_i, x_j) \ln b_{ij}(x_i, x_j) - \sum_i (q_i - 1) \sum_{x_i} b_i(x_i) \ln b_i(x_i)$$

$$U_B = \sum_{(i,j)x_i,x_j} b_{ij}(x_i, x_j) E_{ij}(x_i, x_j) - \sum_i (q_i - 1) \sum_{x_i} b_i(x_i) E_i(x_i)$$

Free Energies

— The Bethe free energy (2)

Then the Bethe free energy is

$$G_B = \sum_{(i,j)x_i,x_j} b_{ij}(x_i, x_j)(E_{ij}(x_i, x_j) + \ln b_{ij}(x_i, x_j))$$

$$- \sum_i (q_i - 1) \sum_{x_i} b_i(x_i)(E_i(x_i) + \ln b_i(x_i))$$

Together with a few normalization and marginalization constraints, the Lagrangian L is as follows:

$$L = G_B + \sum_{(i,j)x_i,x_j} \lambda_{ij}(x_j)[b_j(x_j) - \sum_{x_i} b_{ij}(x_i, x_j)] + \sum_i \beta_i [1 - \sum_{x_i} b_i(x_i)]$$

$$+ \sum_{(i,j)} \sum_{x_i} \lambda_{ji}(x_i)[b_i(x_i) - \sum_{x_j} b_{ij}(x_i, x_j)] + \sum_{(i,j)} \beta_{ij} [1 - \sum_{x_i, x_j} b_{i,j}(x_i, x_j)] \qquad (2)$$

Free Energies

— The Bethe free energy (3)

Taking derivatives of the L wrt the beliefs and those Lagrange multipliers, we have marginal prob. approximation:

$$b_i(x_i) = \frac{1}{Z_i}\exp[-\frac{E_i(x_i)}{T} + \frac{\sum_j \lambda_{ij}(x_i)}{T(q_i - 1)}] \qquad (3)$$

$$b_{ij}(x_i, x_j) = \frac{1}{Z_{ij}}\exp[-\frac{E_{ij}(x_i, x_j)}{T} + \frac{\lambda_{ij}(x_j)}{T} + \frac{\lambda_{ji}(x_i)}{T}] \qquad (4)$$

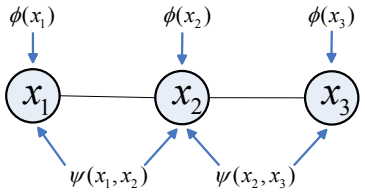
The Bethe approximation is a much better approximation to the exact Gibbs free energy than the mean field approximation. The difficulty lies in the computational part. The Belief Propagation algorithm provides a good solution.

Brief of Belief Propagation(BP) (1)

For pairwise MRF's, the joint prob. distribution for {x} can be factorized

$$p(\{x\}) = \frac{1}{Z}\prod_{i,j}\psi_{ij}(x_i, x_j)\prod_i \phi_i(x_i)$$

where $\psi_{ij}(x_i, x_j)$ tells internal bound between node i and j, and $\phi_i(x_i)$ indicates external evidence at node i.



Brief of Belief Propagation(BP) (2)

Messages 'm' are introduced to pass information between nodes in BP network. The belief (marginal posterior) at a node i is computed as follows:

$$b_i(x_i) = \alpha \phi_i(x_i) \prod_{j \in N(i)} m_{ji}(x_i) \tag{5}$$

and the joint belief (joint marginal posterior) of a pair of neighboring nodes i and j is:

$$b_{ij}(x_i, x_j) = \beta \psi_{ij}(x_i, x_j) \phi_i(x_i) \phi_j(x_j) \prod_{k \in N(i)} m_{ki}(x_i) \prod_{l \in N(j)} m_{lj}(x_j) \tag{6}$$

the message from nodes j to i is:

$$m_{ji}(x_i) \leftarrow \sum_{x_i} \phi_i(x_i) \psi_{ij}(x_i, x_j) \prod_{k \in N(i) \setminus j} m_{ki}(x_i)$$

Equivalence of BP to the Bethe Approximation

By defining

$$\lambda_{ij}(x_j) = T \ln \prod_{k \in N(j) \setminus i} m_{kj}(x_j)$$

it is easily to show that (3) and (4) derived for the Bethe approximation are equivalent to the BP equation (5) and (6).

Equivalence of BP to Dynamic Programming

To get MAP solution (max product) from a belief network, e.g. 3-node graph, the BP algorithm is equivalent to the dynamic programming.

$$\hat{x}_1 = \arg\max_{x_1} \phi(x_1) \max_{x_2} \left[\phi(x_2)\psi(x_1,x_2) \max_{x_3} \left(\phi(x_3)\psi(x_2,x_3) \right) \right]$$

$$\hat{x}_2 = \arg\max_{x_2} \phi(x_2) \left[\max_{x_1} \left(\phi(x_1)\psi(x_1,x_2) \right) \right] \left[\max_{x_3} \left(\phi(x_3)\psi(x_2,x_3) \right) \right]$$

$$\hat{x}_3 = \arg\max_{x_3} \phi(x_3) \max_{x_2} \left[\phi(x_2)\psi(x_2,x_3) \max_{x_1} \left(\phi(x_1)\psi(x_1,x_2) \right) \right]$$

Loopy & non-loopy graph

- BP works for singly connected networks. It is guaranteed to converge to the correct answers.
- BP does not always work for loopy networks. Because same evidence is passed around the network multiple times and mistaken for new evidence.

Loopy graph works sometimes

- Although evidence is “double counted”, all evidence may be double counted. It is proved to be correct in this situation.
- Single loop → BP is guaranteed to generate the most likely state sequence.
- Multiple loops→ Balanced network will work.

BP Visiting Order Reschedule

- In traditional BP algorithm, messages being passed and updated between nodes are without any priority.
- This is not efficient because notes with weak evidence providing less useful information to their neighbors. Messages from these nodes should be passed at later stage compared with those nodes with strong evidence.
- We design a new node visiting order to effectively passing messages between graph nodes.
  1. Rank the nodes according to the belief of their local evidence (breadth first search). Most informative node passes its message first.
  2. Reverse the order in step 1, pass messages back.

Toy Problem 1 - Ising model (1)

- Definition

$$E(\{x\}) = -\sum_{(ij)} J_{ij}(x_i, x_j) - \sum_i h_i(x_i)$$

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

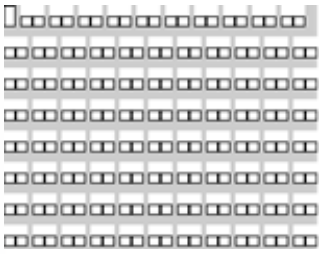
- Specifications

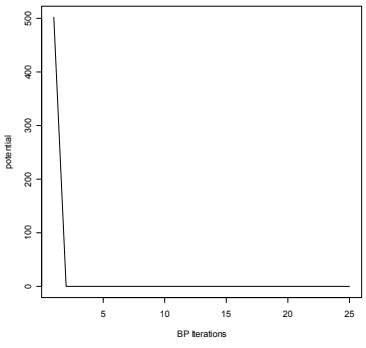
$$J_{ij}(x_i, x_j) = \begin{cases} 1, & x_i \neq x_j \\ 0.1, & x_i = x_j \end{cases}$$

$$h_i(x_i) = 1$$

Toy Problem 1 - Ising model (2)

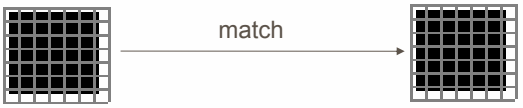
- Bounce back visiting order



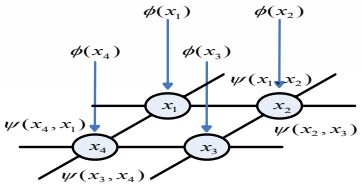


Toy Problem 2 - Rectangle Matching (1)

- Matching black rectangle in two images

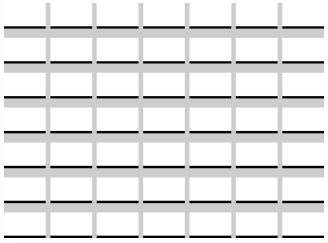


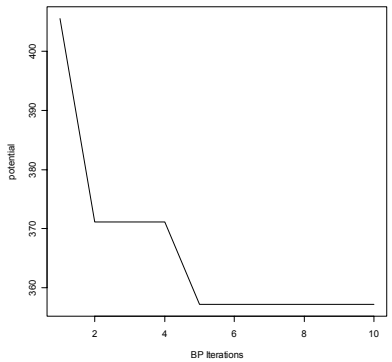
- Build BP graph (7×7 lattice)



Toy Problem 2 - Rectangle Matching (20

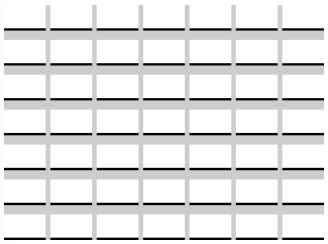
- Ordinary visiting order

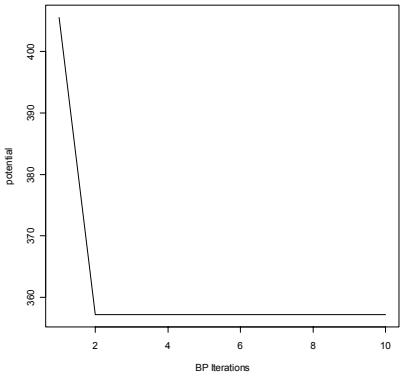




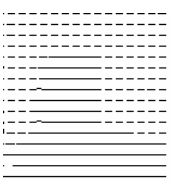
Toy Problem 2 - Rectangle Matching (3)

- Flush visiting order starting from 4 corners



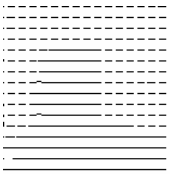


Real Data – Pepsi Can



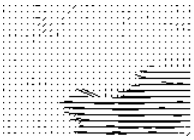
Real Data – Pepsi Can



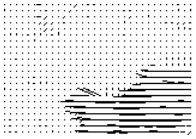

Real Data – Car



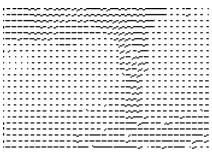

Real Data – Car



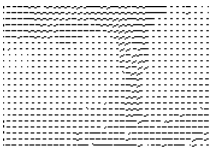

Real Data – Flower Garden



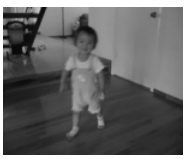
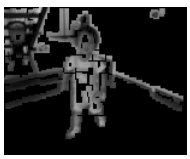



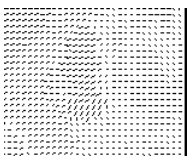

Real Data – Flower Garden



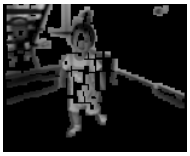



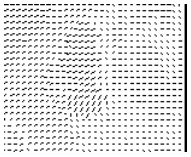

Real Data – Beverly




Real Data – Beverly




Real Data – Printing Room



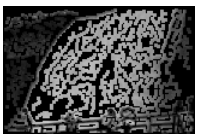


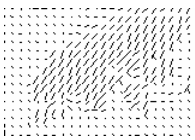
Real Data – Printing Room



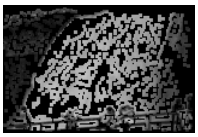


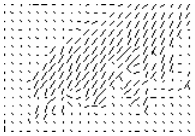
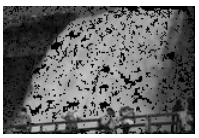
Real Data – Niagara Fall



Real Data – Niagara Fall


Reference

•J. Yedidia, W. T. Freeman and Y. Weiss, *Understanding belief propagation and its generalizations* International Joint Conference on Artificial Intelligence (IJCAI 2001).

•Yedidia, J.S., "An Idiosyncratic Journey Beyond Mean Field Theory", *Advanced Mean Field Methods, Theory and Practice*, ISBN: 0-262-15045-9, pps 21-36, February 2001 .