Belief Propagation and Bethe Free Energy

Seungjin Choi

Department of Computer Science
Pohang University of Science and Technology, Korea
seungjin@postech.ac.kr

Outline

- Helmholtz free energy
- Variational free energy (Gibbs free energy)
- Bethe free energy
- Belief equations on undirected graphs
- Seminar work: J. S. Yedidia, W. T. Freeman, Y. Weiss, "Constructing free energy approximations and generalized belief propagation algorithms," IEEE Trans. Information Theory, vol. 51, no. 7, 2005.

Boltzmann's Law

- Consider a system of N particles, each of which is in one of a discrete number of states, where the states of ith particle are labeled by x_i.
- The overall state of the system is denoted by $\mathbf{x} = [x_1, \ldots, x_N]^\top$.
- Each state of the system has a corresponding energy $E(\mathbf{x})$.
- In thermal equilibrium, the probability of a state is given by Boltzmann's law

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

where T is the temperature which is set as $T = 1$ for the sake of simplicity and Z is the partition function

$$Z = \sum_{\mathbf{x}} e^{-E(\mathbf{x})/T}.$$

Helmholtz Free Energy

- Helmholtz free energy $\mathcal{F}_H$ of a system is given by

$$\boxed{\mathcal{F}_H = -\log Z.}$$

- Now we consider an upper bound on $\mathcal{F}_H$ which is referred to be as Gibb's free energy or variational free energy.

$$\begin{aligned}
\mathcal{F}_H &= \log \left[\sum_{\mathbf{x}} e^{-E(\mathbf{x})} \right] \\
&\leq -\sum_{\mathbf{x}} b(\mathbf{x}) \log \left[\frac{e^{-E(\mathbf{x})}}{b(\mathbf{x})} \right] \\
&= \underbrace{\sum_{\mathbf{x}} b(\mathbf{x}) E(\mathbf{x})}_{U(b)} + \underbrace{\sum_{\mathbf{x}} b(\mathbf{x}) \log b(\mathbf{x})}_{-H(b)} \\
&= \mathcal{F}(b) \quad \text{(variational free energy)}
\end{aligned}$$

Variational Free Energy

Variational free energy $\mathcal{F}(b)$ is defined as

$$\boxed{\mathcal{F}(b) = U(b) - H(b),}$$

where $U(b)$ is the variational average energy and $H(b)$ is the variational entropy

$$\begin{aligned} U(b) &= \sum_{\mathbf{x}} b(\mathbf{x}) E(\mathbf{x}), \\ H(b) &= -\sum_{\mathbf{x}} b(\mathbf{x}) \log b(\mathbf{x}). \end{aligned}$$

Let us compute $\mathcal{F}(b) - \mathcal{F}_H$

$$\begin{aligned} \mathcal{F}(b) - \mathcal{F}_H &= -\sum_{\mathbf{x}} b(\mathbf{x}) \log \left[\frac{e^{-E(\mathbf{x})}}{b(\mathbf{x})} \right] + \log Z \\ &= -\sum_{\mathbf{x}} b(\mathbf{x}) \log \left[\frac{Zp(\mathbf{x})}{b(\mathbf{x})} \right] + \log Z \\ &= \sum_{\mathbf{x}} b(\mathbf{x}) \log \left[\frac{b(\mathbf{x})}{p(\mathbf{x})} \right] \\ &= KL[b \| p]. \end{aligned}$$

Thus, we have

$$\boxed{\mathcal{F}(b) = \mathcal{F}_H + KL[b \| p].}$$

We see that $\mathcal{F}(b) \geq \mathcal{F}_H$ with equality when $b(\mathbf{x}) = p(\mathbf{x})$.

MRF and BP Updates

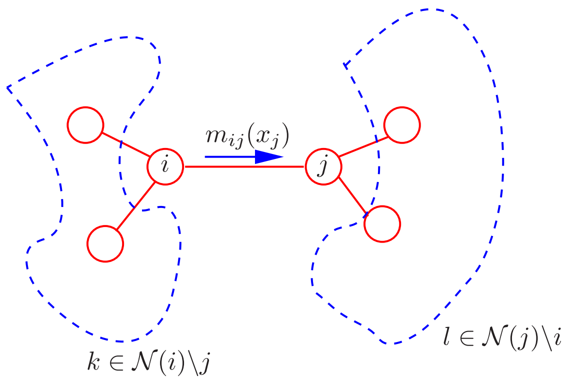
- Consider an undirected graphical model of N nodes with pairwise potentials.
- The joint distribution is given by

$$p(\mathbf{x}) = \frac{1}{Z} \left[\prod_i \prod_j \psi_{ij}(x_i, x_j) \right] \left[\prod_i \psi_i(x_i) \right],$$

  where $\psi_i(x_i)$ is the local evidence for node i and $\psi_{ij}(x_i, x_j)$ is the compatibility matrix between nodes i and j.
- Standard BP update rules are given by

$$\begin{aligned} m_{ij}(x_j) &= \alpha \sum_{x_i} \psi_{ij}(x_i, x_j) \psi_i(x_i) \prod_{k \in \mathcal{N}(i) \backslash j} m_{ki}(x_i), \\ b_i(x_i) &= \alpha \psi_i(x_i) \prod_{k \in \mathcal{N}(i)} m_{ki}(x_i). \end{aligned}$$

Joint Belief



Joint belief $b_{ij}(x_i, x_j)$ is computed as

$$b_{ij}(x_i, x_j) = \alpha \psi_{ij}(x_i, x_j) \psi_i(x_i) \psi_j(x_j) \prod_{k \in \mathcal{N}(i) \backslash j} m_{ki}(x_i) \prod_{l \in \mathcal{N}(j) \backslash i} m_{lj}(x_j).$$

Marginalization

Marginalization condition yields

$$\begin{aligned} b_i(x_i) &= \sum_{x_j} b_{ij}(x_i, x_j) \\ &= \alpha\psi_i(x_i) \prod_{k\in\mathcal{N}(i)\setminus j} m_{ki}(x_i) \underbrace{\sum_{x_j}\left[\psi_{ij}(x_i,x_j)\psi_j(x_j)\prod_{l\in\mathcal{N}(j)\setminus i} m_{lj}(x_j)\right]}_{m_{ji}(x_i)}, \end{aligned}$$

leading to the BP update rule

$$m_{ij}(x_j) = \alpha \sum_{x_i} \psi_{ij}(x_i,x_j)\psi_i(x_i) \prod_{k\in\mathcal{N}(i)\setminus j} m_{ki}(x_i).$$

Mean-Field Approach

- Minimizing the variational free energy $\mathcal{F}(b)$ with respect to a trial probability distribution $b(\mathbf{x})$ is an exact procedure for computing $\mathcal{F}_H$ and recovering $p(\mathbf{x})$.
- As N becomes large, this procedure is intractable as $b(\mathbf{x})$ will take exponentially large memory just to store.
- A practical method is to upper-bound $\mathcal{F}_H$ by minimizing $F(b)$ over a restricted class of probability distributions. (mean-field approach)
- A popular mean-field form for $b(\mathbf{x})$ is the factorized form

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

- Mean-field free energy is of the form

$$\begin{aligned} \mathcal{F}_{MF} &= -\sum_i\sum_j\sum_{x_i}\sum_{x_j} b_i(x_i)b_j(x_j)\log\psi_{ij}(x_i,x_j) \\ &\quad + \sum_i\sum_{x_i} b_i(x_i)\left[\log b_i(x_i) - \log\psi_i(x_i)\right]. \end{aligned}$$

Region-Based Free Energy Approximation

- The factorized mean-field free energy $\mathcal{F}_{MF}$ is a function of single-node beliefs $b_i(x_i)$.
- Cluster variation methods yields the approximate free energy that is a function of beliefs over larger sets of variable nodes.
- Region-based approximations
  - Bethe approximation
  - Kikuchi approximation
- The region-based free energy is given by

$$\mathcal{F}(b_R) = \sum_{r\in R} c_r \left[\sum_{\mathbf{x}_r} b_r(\mathbf{x}_r)E_r(\mathbf{x}_r)\right] + \sum_{r\in R} c_r\left[\sum_{\mathbf{x}_r} b_r(\mathbf{x}_r)\log b_r(\mathbf{x}_r)\right],$$

  where c_r are counting numbers which require $\sum_{r\in R} c_r I_{V_r}(i) = 1$ for $\forall i$, where $I_S(x)$ equals 1 if $x \in S$ and otherwise 0.

Bethe Free Energy

Bethe free energy $\mathcal{F}_{Bethe}$ is given by

$$\begin{aligned} \mathcal{F}_{Bethe} &= \sum_i\sum_j\sum_{x_i}\sum_{x_j} b_{ij}(x_i,x_j)\left[-\log\phi_{ij}(x_i,x_j)\right] \\ &\quad + \sum_i (1-d_i)\sum_{x_i} b_i(x_i)\left[-\log\psi_i(x_i)\right] \\ &\quad + \sum_i\sum_j\sum_{x_i}\sum_{x_j} b_{ij}(x_i,x_j)\log b_{ij}(x_i,x_j) \\ &\quad + \sum_i (1-d_i)\sum_{x_i} b_i(x_i)\log b_i(x_i) \\ &= \sum_i\sum_j\sum_{x_i}\sum_{x_j} b_{ij}(x_i,x_j)\left[\log b_{ij}(x_i,x_j) - \log\phi_{ij}(x_i,x_j)\right] \\ &\quad + \sum_i (1-d_i)\sum_{x_i} b_i(x_i)\left[\log b_i(x_i) - \log\psi_i(x_i)\right], \end{aligned}$$

where $\phi(x_i,x_j) = \psi_{ij}(x_i,x_j)\psi_i(x_i)\psi_j(x_j)$ and d_i is the degree of node i.

Lagrangian

Normalization constraints

$$\sum_{x_i} b_i(x_i) = 1, \qquad \sum_{x_i}\sum_{x_j} b_{ij}(x_i, x_j) = 1.$$

Marginalization constraints

$$\sum_{x_i} b_{ij}(x_i, x_j) = b_j(x_j).$$

Lagrangian is given by

$$\begin{aligned}\mathcal{L} &= \mathcal{F}_{Bethe} + \sum_i \gamma_i \left[1 - \sum_{x_i} b_i(x_i)\right] + \sum_i \sum_j \gamma_{ij} \left[1 - \sum_{x_i}\sum_{x_j} b_{ij}(x_i, x_j)\right] \\ &+ \sum_j \sum_{i \in \mathcal{N}(j)} \sum_{x_j} \lambda_{ij}(x_j) \left[b_j(x_j) - \sum_{x_i} b_{ij}(x_i, x_j)\right].\end{aligned}$$

Main Theorem

Theorem

Let $\{m_{ij}\}$ be a set of BP messages and let $\{b_{ij}, b_i\}$ be the beliefs calculated from those messages. Then the beliefs are fixed points of the BP algorithm if and only if they are zero gradient points of the Bethe free energy $\mathcal{F}_{Bethe}$

$$\begin{aligned}\mathcal{F}_{Bethe} &= \sum_i \sum_j \sum_{x_i} \sum_{x_j} b_{ij}(x_i, x_j) \left[\log b_{ij}(x_i, x_j) - \log \phi_{ij}(x_i, x_j)\right] \\ &+ \sum_i (1 - d_i) \sum_{x_i} b_i(x_i) \left[\log b_i(x_i) - \log \psi_i(x_i)\right],\end{aligned}$$

subject to the normalization and marginalization constraints:

$$\begin{aligned}\sum_{x_i} b_i(x_i) &= 1, \\ \sum_{x_i}\sum_{x_j} b_{ij}(x_i, x_j) &= 1, \\ \sum_{x_i} b_{ij}(x_i, x_j) &= b_j(x_j).\end{aligned}$$

Proof

Solve $\frac{\partial \mathcal{L}}{\partial b_{ij}(x_i, x_j)} = 0$ yields

$$\log b_{ij}(x_i, x_j) = \log \phi_{ij}(x_i, x_j) + \gamma_{ij} - 1 + \lambda_{ij}(x_j) + \lambda_{ji}(x_i).$$

Solve $\frac{\partial \mathcal{L}}{\partial b_i(x_i)} = 0$ yields

$$\log b_i(x_i) = \log \psi_i(x_i) - 1 + \frac{\gamma_i}{1 - d_i} - \frac{1}{1 - d_i} \sum_{j \in \mathcal{N}(i)} \lambda_{ji}(x_i).$$

Setting $\lambda_{ij}(x_j) = \log \prod_{k \in \mathcal{N}(i) \setminus j} m_{ki}(x_i)$ and using the marginalization constraints, one can find that the stationary point conditions on the Lagrangian are equivalent to the BP fixed point conditions.