CS 224 Fall 2023 Scribes: Marvin Li
Nov 20, 2023

Lecture 21: Belief Propagation, Bethe free energy

In this lecture, we describe the Belief Propagation algorithm for variational inference on *Gibbs measures*. We then show that the *Gibbs free energy* on trees can be written in terms of 1- and 2-wise marginals, motivating the notion of *Bethe free energy* which generalizes the idea of Gibbs free energy to a relaxation of probability measures. Finally we prove a novel connection between the fixed points of the Belief Propagation algorithm and the Bethe free energy landscape.

We remind ourselves of our notation for Gibbs measures. Consider an undirected graph $G = (V, E)$ with $V = [n]$ and $E \subset \{(i, j) \in V \times V : i \neq j\}$. One calls a measure μ over the discrete support $\{\pm 1\}^n$ a Gibbs measure if it factors according to the graph, i.e. there exists pairwise *compatibility functions* $\psi_{ij} : \{\pm 1\}^2 \to \mathbb{R}_{\geq 0}$ for each $(i, j) \in E$ such that

$$\mu(x) = \frac{1}{Z} \prod_{(i,j)\in E} \psi_{ij}(x_i, x_j).$$

Z is the *partition function*, the normalization constant of the probability distribution,

$$Z = \sum_{x\in\{\pm 1\}^n} \prod_{(i,j)\in E} \psi_{ij}(x_i, x_j).$$

It is also helpful to define an *energy function* $\mathcal{E}$, where

$$\mathcal{E}(x) = \sum_{(i,j)\in E} \log\left(\frac{1}{\psi_{ij}(x_i, x_j)}\right)$$
$$Z = \sum_{x\in\{\pm 1\}^n} \exp(-\mathcal{E}(x)).$$

In last lecture, we showed that finding a measure $\eta \approx \mu$ is equivalent to minimizing the entropy $H(\eta)$ and maximizing the *average energy* of $\mathcal{E}$ under η.

$$\min_\eta \mathrm{KL}(\nu||\eta) \iff \min_\eta -H(\eta) + \mathbb{E}_{x\sim\eta}[\mathcal{E}(x)] \tag{1}$$

The RHS of Eq. 1 is known as the *Gibbs free energy*. See [MM09],[KZ22],[Mon11] for a thorough introduction.

1 Belief propagation

Belief propagation is a general recipe for variational inference on Gibbs measures. It is provably correct on trees, but practitioners often apply the algorithm to non-trees to great success. First we shall consider the related problem of *marginal estimation*.

1.1 Marginal estimation

Marginal estimation concerns computing quantities like $\mathbb{E}_{x\sim\mu}[x_i]$. Besides variational inference, it is important for Bayesian statistics. Assume we observe y and want to characterize the posterior distribution of some $x|y \sim \mu_y$. The marginal expectations $\{\mathbb{E}_{x\sim\mu_y}[x_i]\}_{i=1}^n$ yield the minimum mean squared error estimator,

$$\begin{aligned}\{\mathbb{E}_{x\sim\mu(y)}[x_i]\}_{i=1}^n &= \text{argmin}_{\hat{x}(\cdot)}\mathbb{E}_{x,y}[\|x-\hat{x}(y)\|^2] \\ &= \mathbb{E}_y[\mathbb{E}_x[\|x-\hat{x}(y)\|^2]].\end{aligned}$$

For any given y, $\mathbb{E}_x[\|x-\hat{x}(y)\|^2]$ is minimized by setting $\hat{x}(y)$ to the posterior means.

1.2 Belief propagation warmup with ISET

To provide additional intuition, we consider an analogue of the belief propagation algorithm for ISET. For $x \in \{\pm 1\}^n$, we define the set $S_x \subset [n]$ s.t. $i \in S_x \iff x_i = 1$. We define μ to be uniform over all independent sets in G with the following compatibility functions,

$$\psi_{ij}(x_i, x_j) = 1[(x_i, x_j) \neq (1,1)].$$

We have $\mu(x) \neq 0$ iff for all $(i,j) \in E$ both x_i, x_j are not 1. For general graphs G, this is #P-complete, but this problem can be solved in polynomial-time with dynamic programming if G is a tree. Let root correspond to the root of the tree, c_v to denote any child $c_v \in \text{children}(v)$, and T_a to denote the subtree rooted at a. We define

$$\begin{aligned}Z_\sigma &:= \text{the number of independent sets of } G \text{ in which the root } x_{\text{root}} = \sigma \\ \mu^{(a)} &:= \text{uniform distribution over independent sets of } T_a \\ Z_\sigma^{(a)} &= \text{the number of independent sets of } T_a \text{ where the root } x_a = \sigma.\end{aligned}$$

Because the intersection of a subtree and an independent set must also be an independent set, we can derive recurrence relationships for Z_+, Z_-. Any independent set U containing the root cannot contain any $\text{children}(\text{root})$. The intersection of U and the subtree $T_{\text{child(root)}}$ must also be an independent set of $T_{\text{child(root)}}$ that does not contain $\text{child}(\text{root})$. When $\{c_1, c_2, \ldots, c_k\} = \text{children}(\text{root})$, then any independent sets $U_1, U_2, \ldots, U_k$ of $T_{c_1}, T_{c_2}, \ldots, T_{c_k}$ that do not contain $c_1, \ldots, c_k$, respectively, form a unique independent set of V that contains the root, $U_1 \cup U_2 \cup \cdots \cup U_k \cup \{\text{root}\}$. Thus, we obtain the recurrence

$$Z_+ = \prod_{c_{\text{root}} \in \text{children(root)}} Z_-^{(c_{\text{root}})}. \tag{2}$$

By similar logic, we can conclude

$$Z_- = \prod_{c_{\text{root}} \in \text{children(root)}} (Z_+^{(c_{\text{root}})} + Z_-^{(c_{\text{root}})}). \tag{3}$$

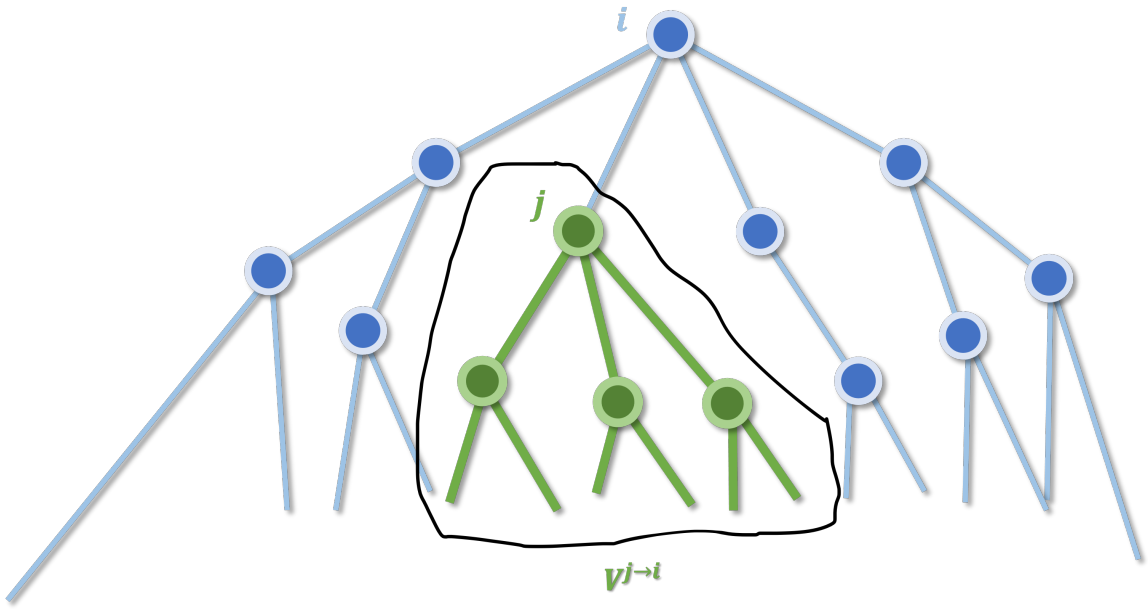
These recurrence relationships provide the basis of a bottom-up dynamic programming algorithm for Z_+, Z_-. For all leaves i, we let $Z_+^{(i)}, Z_-^{(i)} = 1$. We can then use Eqs. 2 and 3 to recursively compute $Z_+^{(i)}, Z_-^{(i)}$ for all $i \in G$. Once we have Z_+, Z_-, we can compute the marginal probability for all nodes in the tree,

$$P_{x\sim\mu}[x_{\text{root}} = +] = \frac{Z_+}{Z_- + Z_+} \propto Z_+.$$

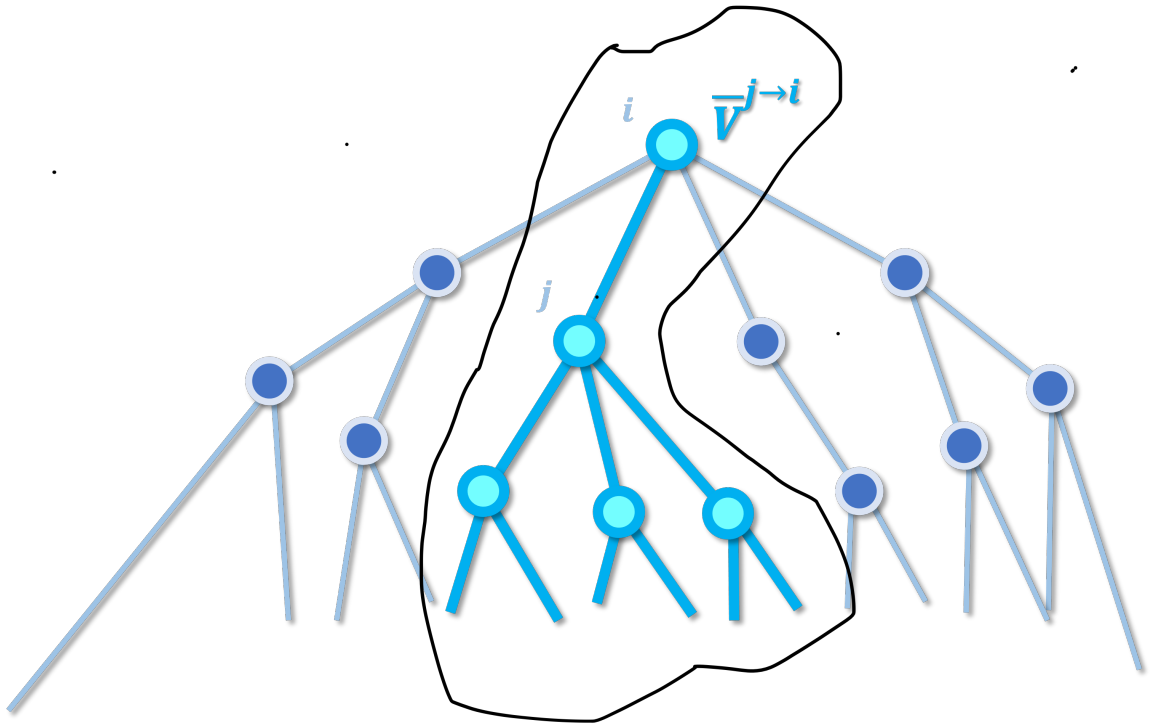
Belief propagation translates the idea of recursing on subtrees to graphs with cycles.

1.3 Belief propagation derivation

We define the subgraph $V^{j\to i}$ to be the subgraph that is still connected to i once we delete the edge (i, j) from G. $V^{j\to i}$ contains all $k \in V$ such that there is a path from j to k that does not cross the edge (i, j) and all edges $(k, \ell) \in E$ such that there is a path from j that includes the edge (k, ℓ) but not the edge (i, j).



It is also useful to define the subgraph $\overline{V}^{j\to i}$, which contains $V_{j\to i}$ as well as i and (i, j).



We can define a Gibbs measure μ with respect to the subgraphs $V^{j\to i}$ or $\overline{V}^{j\to i}$ by including the subset of compatibility functions that corresponds to edges in the subgraph. For notational convenience, we can index the Gibbs measure μ by its subgraph $\mu_{V^{j\to i}}$ or $\mu_{\overline{V}^{j\to i}}$. As in ISET, we want a set of self-similar expressions for each each subgraph $V^{j\to i}, \overline{V}^{j\to i}$ that can then be used to solve for the marginals. The primitive of our self-similar expressions is a *message* from node j to node i,

$$m_\sigma^{j\to i} = \Pr_{x\sim\mu_{V^{j\to i}}}[x_j = \sigma].$$

We can also define a message from node j to i in the subgraph $\overline{V}^{j\to i}$,

$$\overline{m}_\sigma^{j\to i} = \Pr_{x\sim\mu_{\overline{V}^{j\to i}}}[x_i = \sigma].$$

Note that we can express Z^+ and Z^- from the ISET example in terms of the messages,

$$Z_+ \propto \prod_{j\in\partial i} m_-^{j\to i}$$

$$Z_- \propto \prod_{j\in\partial i} (m_+^{j\to i} + m_-^{j\to i}).$$

Crucially, we can also compute the marginal probabilities in terms of the messages. By the law of total probability, we can express $\Pr_{x\sim\mu}[x_i = \sigma]$ as a sum of the probability over all $s \in \{\pm 1\}^n$ where $s_i = \sigma$. We use the shorthand s_K to denote indexing s by the elements of K and s_{-K} to denote indexing s by the elements of $[n] - K$.

$$\Pr_{x\sim\mu}[x_i = \sigma] \propto \sum_{s_i=\sigma; s_{-i}\in\{\pm1\}^{n-1}} \prod_{(i,j)\in E} \psi_{ij}(x_i, x_j) \tag{4}$$

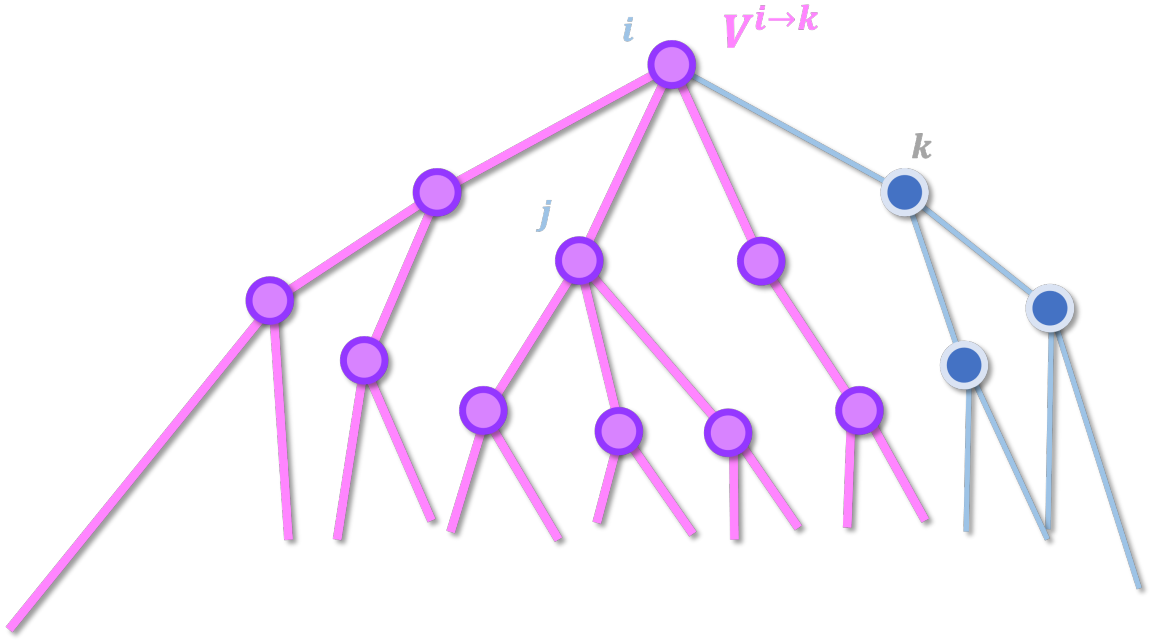
$$= \sum_{s_i=\sigma; s_{-i}\in\{\pm1\}^{n-1}} \prod_{j\in\partial i} \left[\psi_{ij}(x_i, x_j) \prod_{(k,\ell)\in E^{j\to i}} \psi_{k\ell}(x_k, x_\ell) \right] \tag{5}$$

$$= \prod_{j\in\partial i} \left[\sum_{s_{V_{j\to i}}\in\{\pm1\}^{|V_{j\to i}|}} \psi_{ij}(x_i, x_j) \prod_{(k,\ell)\in E^{j\to i}} \psi_{k\ell}(x_k, x_\ell) \right] \tag{6}$$

$$= \prod_{j\in\partial i} \left[\sum_{s_j\in\{\pm1\}} \psi_{ij}(x_i, x_j) \sum_{s_{V_{j\to i}-\{j\}}\in\{\pm1\}^{|V_{j\to i}|-1}} \prod_{(k,\ell)\in E^{j\to i}} \psi_{k\ell}(x_k, x_\ell) \right] \tag{7}$$

$$\propto \prod_{j\in\partial i} \sum_{s\in\{\pm1\}} \psi_{ij}(x_i, x_j) m_s^{\textcircled{j}\to i}. \tag{8}$$

The factor of proportionality is the same for $\sigma = \pm 1$, so we can simply normalize the RHS of 8 to compute the marginal probabilities. We can also relate the marginal probabilities to the messages. When G is a tree, $V^{i\to k}$ is the graph G excluding the subtree under k.



This means the message of $V^{i\to k}$ is equal to

$$m_\sigma^{\textcircled{i}\to k} = \Pr_{x\sim\{i\}\cup\bigcup_{j\in\partial i\setminus\{k\}} T_j}[x_i = \mu] \propto \prod_{j\in\partial i\setminus\{k\}} \sum_{s\in\{\pm 1\}} \psi_{ij}(x_i, x_j) m_s^{\textcircled{j}\to i}. \tag{9}$$

From Eq. 8, we know $\sum_{s\in\{\pm1\}} \psi_{ij}(x_i, x_j) m_s^{\textcircled{j}\to i}$ is proportional to $\overline{m}_\sigma^{j\to\textcircled{i}}$. We have the following recursive equations,

$$m_\sigma^{\textcircled{i}\to k} \propto \prod_{j\in\partial i\setminus\{k\}} \overline{m}_\sigma^{j\to\textcircled{i}} \tag{10}$$

$$\overline{m}_\sigma^{j\to\textcircled{i}} \propto \sum_{s\in\{\pm1\}} \psi_{ij}(\sigma, s) m_s^{\textcircled{j}\to i} \tag{11}$$

$$\Pr_{x\sim\mu}[x_i = \sigma] \propto \prod_{j\in\partial i} \overline{m}_\sigma^{i\to\textcircled{j}} \propto m_\sigma^{\textcircled{i}\to k} \cdot \overline{m}^{k\to\textcircled{i}} \tag{12}$$

For trees, we can compute the messages with a bottom-up dynamic programming algorithm. We pick an arbitrary vertex as the root and assign $m_\sigma^{\textcircled{i}\to j} = \frac{1}{2}$ for all leafs i and parents j. Eqs. 9 and 11 allow us to recursively compute $m_\sigma^{\textcircled{i}\to j}, \overline{m}_\sigma^{j\to\textcircled{i}}$ for all edges (i, j). Then we can apply Eq. 12 to compute the marginal probabilities.

Belief Propagation can also be adapted for more general graphs. Instead of recursing, we iteratively update $\{m_\sigma^{\textcircled{i}\to j}, \overline{m}_\sigma^{i\to\textcircled{j}}\}_{(i,j)\in E,\sigma\in\{\pm1\}}$ according to Eqs. 10

and 11 until convergence. Because the updating step can be done in parallel over all messages, this is very fast and efficient algorithm in practice.

Belief Propagation is extremely hard to analyze rigorously on the non-acyclic graphs. Koehler 2019 proved that message-passing algorithms, e.g. Belief Propagation, find the global minimum of Bethe free energy for ferromagnetic Ising models, a special class of Gibbs measures, on any graph [Koe19]. Belief Propagation may also work more generally for sparse, random graphs because they "locally" resemble a tree. Consider a random graph where each edge is included with probability c/n and a vertex v. There are $\approx c$ neighbors of v or its descendants and thus $\approx c^d$ descendants of v within d edges of v. Because each child is equally likely to be one of n nodes, the probability v does not have a path to itself of length $\leq d$ is $\approx (1-1/n)^{c^d} \approx e^{-c^d/n}$. When $c^d << n$, i.e. the graph is sparse, this is unlikely.

1.4 Higher-order marginals

Belief Propagation can also be used to compute the 2-wise marginals when G is a tree. Let $(i_1, i_2) \in E$ and $\partial(i_1, i_2) = \partial i_1 \cup \partial i_2 - \{i_1, i_2\}$. For $j \in \partial(i_1, i_2)$, we let $i(j)$ denote the unique neighbor. Using the Law of Total Probability and Markov's property, we can marginalize over all other vertices $V \backslash \partial(i_1, i_2)$.

$$\begin{aligned} Pr_{x\sim\mu}[(x_{i_1}, x_{i_2}) = (\sigma_{i_1}, \sigma_{i_2})] &\propto \psi_{i_1 i_2}(\sigma_{i_1}, \sigma_{i_2}) \sum_{s\in\{\pm 1\}^{\partial(i_1,i_2)}} \prod_{j\in\partial(i_1,i_2)} P_{V^{j\to i(j)}}[x_j = s_j]\psi_{i(j)j}(\sigma_{i(j)}, s_j) \\ &\propto \psi_{i_1 i_2}(\sigma_{i_1}, \sigma_{i_2}) \prod_{j\in\partial(i_1,i_2)} \sum_{s_j\in\{\pm 1\}} P_{V^{j\to i(j)}}[x_j = s_j]\psi_{i(j)j}(\sigma_{i(j)}, s_j) \\ &\propto \psi_{i_1 i_2}(\sigma_{i_1}, \sigma_{i_2}) \prod_{j\in\partial(i_1,i_2)} \sum_{s_j\in\{\pm 1\}} \overline{m}^{j\to i(j)}_{\sigma_{i(j)}}. \end{aligned}$$

Interestingly, we can actually express the Gibbs free energy on trees in terms of 1 and 2-wise marginals. The average energy of a measure μ is

$$\mathbb{E}_\mu[\mathcal{E}(x)] = \sum_{(i,j)\in E} \mathbb{E}_\mu[\log 1/\psi_{ij}(x_i, x_j)].$$

Note that $\mathbb{E}_\mu[\log 1/\psi_{ij}(x_i, x_j)]$ only depends on the pairwise marginals. The derivation that entropy can be expressed in 1- and 2-wise marginals is more involved.

Lemma 1. *Let $\mu(x_i)$ and $\mu_{ij}(x_i, x_j)$ denote the 1- and 2-wise marginals. If G is a tree, then*

$$\mu(x) = \prod_{(i,j)\in E} \mu_{ij}(x_i, x_j) \prod_{i\in[n]} \mu_i(x_i)^{1-|\partial i|}$$

Moreover, the entropy of G is a function of $\mu(x_i)$ and $\mu_{ij}(x_i, x_j)$.

Proof. We prove by induction. The base case ($n = 1$) is trivially true. When $n > 1$, let i be any leaf of the tree connected by (i, j). We find $\Pr_\mu[x = s]$ is equal to

$$\begin{aligned}\Pr_\mu[x = s] &= \Pr_\mu[x_{[n]\setminus\{i\}} = s_{[n]\setminus\{i\}}] \Pr_\mu[x_i = s_i | x_{[n]\setminus\{i\}} = s_{[n]\setminus\{i\}}] \\ &= \Pr_\mu[x_{[n]\setminus\{i\}} = s_{[n]\setminus\{i\}}] \Pr_\mu[x_i = s_i | x_j = s_j] \\ &= \Pr_\mu[x_{[n]\setminus\{i\}} = s_{[n]\setminus\{i\}}] \frac{\mu_{ij}(x_i, x_j)}{\mu_j(x_j)}.\end{aligned}$$

We can view $\Pr_\mu[x_{[n]\setminus\{i\}} = s_{[n]\setminus\{i\}}]$ as the Gibbs measure of the subgraph $V\setminus\{i\}$, the tree where we remove (i, j) and i. By our inductive hypothesis, we can express this in terms of the 1- and 2-wise marginals,

$$Pr_\mu[x_{[n]\setminus\{i\}} = s_{[n]\setminus\{i\}}] = \mu_{V\setminus\{i\}}(x) = \prod_{(k,\ell)\in E\setminus\{(i,j)\}} \mu_{k\ell}(x_k, x_\ell) \prod_{k\in[n]\setminus\{i\}} \mu_k(x_k)^{1-|\partial_{V\setminus\{i\}}k|}.$$

We obtain

$$\begin{aligned}\Pr_\mu[x = s] &= \left[\prod_{(k,\ell)\in E\setminus\{(i,j)\}} \mu_{k\ell}(x_k, x_\ell) \prod_{k\in[n]\setminus\{i\}} \mu_k(x_k)^{1-|\partial_{V\setminus\{i\}}k|}\right] \frac{\mu_{ij}(x_i, x_j)}{\mu_j(x_j)} \\ &= \prod_{(k,\ell)\in E} \mu_{k\ell}(x_k, x_\ell) \prod_{k\in[n]} \mu_k(x_k)^{1-|\partial k|}.\end{aligned}$$

Furthermore, we can calculate the entropy of μ in terms of the 1- and 2-wise marginals,

$$\begin{aligned}H(\mu) &= -\mathbb{E}_{x\sim\mu}[\log \Pr_\mu[x]] \\ &= -\mathbb{E}_{x\sim\mu}\left[\log \mu_{k\ell}(x_k, x_\ell) + \sum_{k\in[n]} (1 - |\partial k|) \log \mu_k(x_k)\right] \\ &= -\sum_{(k,\ell)\in E} \mathbb{E}_{x\sim\mu}\left[\log \mu_{k\ell}(x_k, x_\ell)\right] - \sum_{k\in[n]} (1 - |\partial k|)\mathbb{E}_{x\sim\mu}\left[\log \mu_k(x_k)\right] \\ &= \sum_{(k,\ell)\in E} H(\mu_{k\ell}(x_k, x_\ell)) + \sum_{k\in[n]} (1 - |\partial k|)H(\mu_k(x_k)).\end{aligned}$$

□

If you are more curious about Belief Propagation, see [KE22].

2 Bethe Free Energy

Bethe Free energy is a generalization of Gibbs free energy to a certain class of pseudo-distributions. More precisely, we can define Bethe Free Energy over marginals

$\{\nu_i, \nu_{ij}\}$ that satisfy a *local consistency* property. Every valid Gibbs measure defines a locally consistent set of marginals, but there may not exist a distribution for every locally consistent set of marginals. The *local consistency* property simply states that marginalizing over 2-wise marginals obtains the 1-wise marginals.

$$\sum_{x_j \in \{\pm 1\}} \nu_{ij}(x_i, x_j) = \nu_i(x_i) \quad \forall (i,j) \in E, x_i \pm \{\pm 1\}.$$

These are also known as degree-2 Sherali-Adams constraints. Now we are ready to define the *Bethe Free Energy* in terms of 1- and 2-wise marginals,

$$G_\beta[\nu] := -\sum_{(k,\ell)\in E} H(\nu_{k\ell}(x_k, x_\ell)) + \sum_{k\in[n]} (|\partial k| - 1) H(\nu_k(x_k)) + \mathbb{E}_\nu[\mathcal{E}].$$

Bethe free energy is a functional over the space of pseudo-distributions. We have already seen an equivalence between Bethe Free Energy and Gibbs Free Energy in trees. Because Belief Propagation is exact on trees, one could hope for a similar equivalence between Belief Propagation and Bethe Free Energy on more general types of graphs. In fact, we will show that the fixed points of Belief Propagation are equivalent to the stationary points of Bethe Free Energy on all types of graphs.

Lemma 2. *Fixed points of the Belief Propagation algorithm satisfy local consistency.*

Proof. Belief Propagation yields the following marginals,

$$\nu_i(\sigma) = \frac{1}{Z_i} \prod_{j\in\partial i} \overline{m}_\sigma^{j\to \textcircled{i}}$$

$$\nu_{ij}(\sigma_i, \sigma_j) = \frac{1}{Z_{ij}} \psi_{ij}(\sigma_i, \sigma_j) m_{\sigma_i}^{\textcircled{i}\to j} m_{\sigma_j}^{\textcircled{j}\to i},$$

with the normalizing constants Z_i, Z_{ij},

$$Z_i = \sum_{\sigma\in\{\pm1\}} \prod_{j\in\partial i} \overline{m}_\sigma^{j\to i}$$

$$Z_{ij} = \sum_{\sigma_i,\sigma_j\in\{\pm1\}} \psi_{ij}(\sigma_i, \sigma_j) m_{\sigma_i}^{\textcircled{i}\to j} m_{\sigma_j}^{\textcircled{j}\to i}.$$

For any $x_i \in \{\pm 1\}$, we have

$$\sum_{x_j\in\{\pm1\}} \nu_{ij}(x_i, x_j) = \frac{1}{Z_{ij}} \sum_{x_j\in\{\pm1\}} \psi_{ij}(x_i, x_j) m_{x_i}^{\textcircled{i}\to j} m_{x_j}^{\textcircled{j}\to i} \tag{13}$$

$$= \frac{m_{x_i}^{\textcircled{i}\to j}}{Z_{ij}} \sum_{x_j\in\{\pm1\}} \psi_{ij}(x_i, x_j) m_{x_j}^{\textcircled{j}\to i} \tag{14}$$

$$\propto m_{x_i}^{\textcircled{i}\to j} \overline{m}_{x_i}^{\textcircled{i}\to i} \tag{15}$$

$$\propto \nu(x_i). \tag{16}$$

where Eq. 14 applies the fixed point condition. The LHS and RHS marginalize to 1 when we sum over $x_i = \pm 1$, so the proportionality is actually equality. □

We can also rewrite Bethe free energy in terms of the partitions.

Lemma 3. *We also define*

$$Z_{i;j} := \sum_{\sigma \in \{\pm 1\}} m_\sigma^{ⓘ \to j} \overline{m}_\sigma^{j \to ⓘ}.$$

Then

$$G_\beta[\nu] = -\sum_{i \in [n]} \log Z_i - \sum_{(i,j) \in E} \log Z_{ij} + \sum_{i \in [n], j \in \partial i} \log Z_{i;j}.$$

Proof. We can compute the entropy of the 2-wise marginals from our definition of ν,

$$\begin{aligned}
-\sum_{(i,j)\in E} H(\nu_{ij}) + \mathbb{E}_\nu[\mathcal{E}] &= -\sum_{(i,j)\in E} \mathbb{E}_{\nu_{ij}} \log \frac{1}{\nu_{ij}(x_i, x_j)} - \sum_{(i,j)\in E} \mathbb{E}_{\nu_{ij}} \log \frac{1}{\psi_{ij}(x_i, x_j)} \\
&= -\sum_{(i,j)\in E} \mathbb{E}_{\nu_{ij}} \log \frac{\psi_{ij}(x_i, x_j)}{\nu_{ij}(x_i, x_j)} \\
&= -\sum_{(i,j)\in E} \mathbb{E}_{\nu_{ij}} \log \frac{Z_{ij}}{m_{x_i}^{ⓘ \to j} m_{x_j}^{ⓙ \to i}} \\
&= -\sum_{(i,j)\in E} \log Z_{ij} + \sum_{(i,j)\in E} \mathbb{E}_{\nu_{ij}} [\log m_{x_i}^{ⓘ \to j} + \log m_{x_j}^{ⓙ \to i}].
\end{aligned}$$

We can also compute the entropy of the 1-wise marginals from our definition of ν,

$$\begin{aligned}
-\sum_{i \in [n]} H(\nu_i) &= -\sum_i \mathbb{E}_{\nu_i} \log \frac{1}{\nu_i(x_i)} \\
&= -\sum_i \mathbb{E}_{\nu_i} \log \frac{Z_i}{\prod_{j \partial i} \overline{m}_{x_i}^{j \to ⓘ}} \\
&= -\sum_i \log Z_i + \sum_i \mathbb{E}_{\nu_i} \left[\sum_j \log \overline{m}_{x_i}^{j \to ⓘ} \right].
\end{aligned}$$

It will be useful to compute $\sum_{i\in[n]} |\partial i| H(\nu_i)$ using the Eq. 12 fixed point condition,

$$\begin{aligned}
\sum_{i\in[n]} |\partial i| H(\nu_i) &= \sum_i \sum_{j\in\partial i} \mathbb{E}_{\nu_i} \log \frac{1}{\nu_i(x_i)} \\
&= \sum_i \sum_{j\in\partial i} \mathbb{E}_{\nu_i} \log \frac{Z_{i;j}}{m_{x_i}^{\textcircled{i}\to j} \overline{m}_{x_i}^{j\to\textcircled{i}}} \\
&= \sum_i \sum_{j\in\partial i} \log Z_{i;j} - \sum_i \sum_{j\in\partial i} \mathbb{E}_{\nu_{ij}} [\log m_{x_i}^{\textcircled{i}\to j} + \log \overline{m}_{x_i}^{j\to\textcircled{i}}].
\end{aligned}$$

Finally, we can compute

$$\begin{aligned}
G_\beta[\nu] &= -\sum_{(k,\ell)\in E} H(\nu_{k\ell}(x_k, x_\ell)) + \sum_{k\in[n]} (|\partial k| - 1) H(\nu_k(x_k)) + \mathbb{E}_\nu[\mathcal{E}] \\
&= \left[-\sum_{(i,j)\in E} \log Z_{ij} + \sum_{(i,j)\in E} \mathbb{E}_{\nu_{ij}} [\log m_{x_i}^{\textcircled{i}\to j} + \log m_{x_j}^{\textcircled{j}\to i}] \right] \\
&+ \left[\sum_i \sum_{j\in\partial i} \log Z_{i;j} - \sum_i \sum_{j\in\partial i} \mathbb{E}_{\nu_{ij}} [\log m_{x_i}^{\textcircled{i}\to j} + \overline{m}_{x_i}^{j\to\textcircled{i}}] \right] \\
&- \left[\sum_i \log Z_i + \sum_i \mathbb{E}_{\nu_i} \left[\sum_j \log \overline{m}_{x_i}^{j\to\textcircled{i}} \right] \right] \\
&= -\sum_{i\in[n]} \log Z_i - \sum_{(i,j)\in E} \log Z_{ij} + \sum_{i\in[n], j\in\partial i} \log Z_{i;j}.
\end{aligned}$$

□

Note that $Z_i, Z_{i;j}, Z_{ij}$ are invariant to the scaling of the messages, so we do not need to normalize them to compute Bethe Free Energy. In general, there may be many fixed points of the belief propagation algorithm that are not the correct marginals. However, these fixed points are connected to Bethe Free Energy by the following theorem.

Theorem 1. *Take any graph $G = (V, E)$, which may not be a tree. There is a 1-1 correpondence between messages that are*

$$\textit{fixed points of Belief Propagation} \iff \textit{stationary points for } G_\beta.$$

Proof. We compute the stationary points of Bethe free energy by differentiating with

respect to the messages and ignoring irrelevant terms that do not depend on $m_{\sigma_i}^{\textcircled{i}\to j}$,

$$\begin{aligned}\frac{\partial G_\beta[\nu]}{\partial m_{\sigma_i}^{\textcircled{i}\to j}} &= \frac{1}{Z_{i;j}}\frac{\partial Z_{i;j}}{\partial m_{\sigma_i}^{\textcircled{i}\to j}} - \frac{1}{Z_{ij}}\frac{\partial}{\partial m_{\sigma_i}^{\textcircled{i}\to j}} Z_{ij} \\ &= \frac{\overline{m}_{\sigma_i}^{j\to\textcircled{i}}}{\sum_{\sigma\in\{\pm1\}} m_\sigma^{\textcircled{i}\to j}\overline{m}_\sigma^{j\to\textcircled{i}}} - \frac{\sum_{\sigma_j\in\{\pm1\}}\psi_{ij}(\sigma_i,\sigma_j)m_{\sigma_j}^{\textcircled{j}\to i}}{\sum_{\sigma,\sigma_j\in\{\pm1\}} m_\sigma^{\textcircled{i}\to j} m_\sigma^{\textcircled{j}\to i}}.\end{aligned}$$

This vanishes for all i, j, σ_i iff

$$\overline{m}_{\sigma_i}^{j\to\textcircled{i}} \propto \sum_{\sigma_j}\psi_{ij}(\sigma_i,\sigma_j)m_{\sigma_j}^{\textcircled{j}\to i}.$$

This is exactly Eq. 11 from the Belief Propagation algorithm! We can also find Eq.10 in the Belief Propagation algorithm by differentiating with respect to $\overline{m}_{\sigma_i}^{j\to\textcircled{i}}$,

$$\begin{aligned}\frac{\partial G_\beta[\nu]}{\partial \overline{m}_{\sigma_i}^{j\to\textcircled{i}}} &= \frac{1}{Z_{i;j}}\frac{\partial}{\partial \overline{m}_{\sigma_i}^{j\to circledi}} Z_{i;j} - \frac{1}{Z_i}\frac{\partial}{\partial \overline{m}_\sigma^{j\to\textcircled{i}}} Z_i \\ &= \frac{m_{\sigma_i}^{\textcircled{i}\to j}}{\sum_{\sigma\in\{\pm1\}} m_\sigma^{\textcircled{i}\to j}\overline{m}_\sigma^{j\to\textcircled{i}}} - \frac{\prod_{k\in\partial i\setminus\{j\}}\overline{m}_{\sigma_i}^{k\to\textcircled{i}}}{\sum_\sigma\prod_{k\in\partial i}\overline{m}_\sigma^{k\to\textcircled{i}}}.\end{aligned}$$

This equals 0 for all i, j, σ_i iff

$$m_{\sigma_i}^{\textcircled{i}\to j} \propto \prod_{k\in\partial i\setminus\{j\}}\overline{m}_{\sigma_i}^{k\to\textcircled{i}}.$$

□

For a more in-depth discussion of Bethe Free Energy, see [Pfi14],[Mac11].

References

[KE22] Volodymyr Kuleshov and Stefano Ermon. Cs 228 - probabilistic graphical models. chapter 7. 2022.

[Koe19] Frederic Koehler. Fast convergence of belief propagation to global optima: Beyond correlation decay. *CoRR*, abs/1905.09992, 2019.

[KZ22] Florent Krzakala and Lenka Zdeborová. Statisticalphysics methods in optimization and machine learning an introduction to replica, cavity message-passing technique. chapter 4. 2022.

[Mac11] Nicolas Macris. Lecture notes 10: Bethe free energy and its relation to bp in statistical physics for communication and computer science. 2011.

[MM09] Marc Mezard and Andrea Montanari. Information, physics, and computation. chapter 14. Oxford University Press, Inc., USA, 2009.

[Mon11] Andrea Montanari. Lecture notes for stat 375 inference in graphical models. chapter 2. 2011.

[Pfi14] Henry D. Pfister. The gibbs free energy, the bethe free entropy, and the sum-product algorithm. 2014.