The Gibbs Free Energy, the Bethe Free Entropy, and the Sum-Product Algorithm

Supplemental Material for Graphical Models and Inference
Henry D. Pfister

November 5th, 2014

1 Introduction

Let $\mathcal{X}$ be a finite set and $f : \mathcal{X}^n \to \mathbb{R}_{\geq 0}$ be a non-negative real function. For a finite set $\mathcal{A}$, let $\mathcal{P}(\mathcal{A})$ be the set of probability distributions over $\mathcal{A}$. The function f implicitly defines a probability distribution $\mu \in \mathcal{P}(\mathcal{X}^n)$ via

$$\mu(\underline{x}) \triangleq \frac{1}{Z} f(\underline{x}),$$

where $Z = \sum_{\underline{x} \in \mathcal{X}^n} f(\underline{x})$ is called the partition function (or sum). For $[n] \triangleq \{1, 2, \ldots, n\}$ and a subset $I = \{i_1, i_2, \ldots, i_k\} \subseteq [n]$ with $i_1 < i_2 < \cdots < i_k$, let $\underline{x}_I \triangleq (x_{i_1}, x_{i_2}, \ldots, x_{i_k})$ denote the ordered subvector. For $\mu \in \mathcal{P}(\mathcal{X}^n)$, the marginal distribution of $\underline{x}_I$ is denoted

$$\mu_I(\underline{x}_I) \triangleq \sum_{\underline{x}_{[n]\setminus I}} \mu(\underline{x})$$

or in shorthand as $\mu(\underline{x}_I)$. Likewise, the conditional distribution of $\underline{x}_I$ given $\underline{x}_J$ is denoted

$$\mu_{I|J}(\underline{x}_I | \underline{x}_J) \triangleq \frac{\mu_{I\cup J}(\underline{x}_{I\cup J})}{\mu_J(\underline{x}_J)}$$

or in shorthand as $\mu(\underline{x}_I | \underline{x}_J)$. The Shannon entropy (in nats) of any $\nu \in \mathcal{P}(\mathcal{A})$ is denoted

$$H(\nu) \triangleq \sum_{a \in \mathcal{A}} \nu(a) \ln \frac{1}{\nu(a)}.$$

2 The Gibbs Free Energy

In physics, the configuration distribution is denoted μ_β and the partition function is denoted $Z(\beta)$ due to their dependence on the inverse temperature parameter β. In this case, the quantity $-\frac{1}{\beta} \ln Z(\beta)$ is called the *free energy* of the system and the quantity $\ln Z(\beta)$ is called the *free entropy* of the system. The choice of $\beta = 1$ is natural for inference problems and, thus, the definitions of μ and Z above assume $\beta = 1$. Therefore, one finds that the free entropy is simply the *log partition-function* $\ln Z$ and the free energy is simply the *negative log partition-function* $-\ln Z$.

Remark 2.1. If $f(\underline{x}) \in \{0, 1\}$, then $\mu(\underline{x})$ is uniform over the set of valid patterns $\mathcal{V} = \{\underline{x} \in \mathcal{X}^n \mid f(\underline{x}) = 1\}$ and $Z = |\mathcal{V}|$ is the number of valid patterns. In this case, the free entropy $\ln Z$ equals the Shannon entropy $H(\mu)$.

The Gibbs free energy $G : \mathcal{P}(\mathcal{X}^n) \to \mathbb{R}$ maps $\nu \in \mathcal{P}(\mathcal{X}^n)$ to the real numbers

$$G(\nu) = \underbrace{\sum_{\underline{x} \in \mathcal{X}^n} \nu(\underline{x}) \ln \frac{1}{f(\underline{x})}}_{U(\nu)} - H(\nu), \tag{1}$$

where $U(\nu)$ is the average "energy" of the system for the configuration distribution ν. Since $f(\underline{x}) = Z\mu(\underline{x})$, one can also write the Gibbs free energy as

$$\begin{aligned} G(\nu) &= \sum_{\underline{x}\in\mathcal{X}^n} \nu(\underline{x}) \ln \frac{1}{Z\mu(\underline{x})} - \sum_{\underline{x}\in\mathcal{X}^n} \nu(\underline{x}) \ln \frac{1}{\nu(\underline{x})} \\ &= -\ln Z + \underbrace{\sum_{\underline{x}\in\mathcal{X}^n} \nu(\underline{x}) \ln \frac{\nu(\underline{x})}{\mu(\underline{x})}}_{D(\nu||\mu)}, \end{aligned} \tag{2}$$

where $D(\nu||\mu)$ is the Kullback-Liebler divergence. Since $D(\nu||\mu) \geq 0$ with equality iff $\nu = \mu$, it follows that $\min_{\nu\in\mathcal{P}(\mathcal{X})} G(\nu) = -\ln Z$ is achieved uniquely at $\nu^* = \mu$. This formulation is often called the *variational Gibbs free energy*.

To motivate this definition, we mention a connection to the theory of large deviations. Let P_N be the probability that the empirical distribution of N samples drawn according to μ is equal to ν. If $N\nu(\underline{x})$ is integer for all $\underline{x} \in \mathcal{X}^n$, then [1, Thm. 11.1.4] shows that $P_N = e^{-N[D(\nu||\mu)+o(1)]}$ and, thus, $D(\nu||\mu) = G(\nu) + \ln Z$ implies that

$$P_N = e^{-N[G(\nu)+\ln Z+o(1)]}.$$

3 Factor Graphs

Many operations on f can be simplified if f factors into the product of local potentials. These simplified operations can often be understood in terms of a factor graph. A *factor graph* is bipartite graph defined by a set of variable nodes $V = [n]$, a set of factor nodes F, a set of edges $E \subseteq V \times F$, and a weight function f_a for each factor $a \in F$. Let $V(a) \triangleq \{i \in V \mid (i,a) \in E\}$ denote the set of variable nodes adjacent to a and $F(i) \triangleq \{a \in F \mid (i,a) \in E\}$ denote the set of factor nodes adjacent to i. Such a factor graph represents the factorization

$$f(x_1,\ldots,x_n) = \left(\prod_{a\in F} f_a(\underline{x}_{V(a)})\right)\left(\prod_{i\in V} f_i(x_i)\right).$$

From this, we see that a variable node $i \in V$ participates in factor $a \in F$ iff $i \in V(a)$. Likewise, a factor $a \in F$ depends on variable $i \in V$ iff $a \in F(i)$.

3.1 Factor Graphs without Cycles

If the factor graph does not have any cycles, then inference and analysis are both greatly simplified. In particular, the sum-product algorithm (SPA), which is also called belief propagation (BP), can be used to efficiently compute the *factor marginals* $\{\mu_{V(a)}\}_{a\in F}$ and $\{\mu_i\}_{i\in V}$.

The message-passing update rules of the SPA are given by

$$\begin{aligned} \mu^{(\ell+1)}_{j\to a}(x) &\propto f_j(x) \prod_{b\in F(j)\setminus a} \hat{\mu}^{(\ell)}_{b\to j}(x) \\ \hat{\mu}^{(\ell)}_{a\to j}(x) &\propto \sum_{\underline{x}_{V(a)}} f_a(\underline{x}_{V(a)})\delta_{x_j,x} \prod_{i\in V(a)\setminus j} \mu^{(\ell)}_{i\to a}(x_i), \end{aligned} \tag{3}$$

along with the normalization conditions $\sum_{x\in\mathcal{X}} \mu^{(\ell)}_{j\to a}(x) = 1$ and $\sum_{x\in\mathcal{X}} \hat{\mu}^{(\ell)}_{a\to j}(x) = 1$. The symbol $\delta_{x_j,x}$ denotes the Kronecker delta function and equals 1 if $x_j = x$ and 0 otherwise. The algorithm is typically initialized to $\mu^{(0)}_{j\to a}(x) \propto f_j(x)$. If the factor graph does not have cycles, then this iteration converges to a fixed point after a finite number of steps and we denote the fixed point messages by $\hat{\mu}^*_{a\to j}(x)$ and $\mu^*_{j\to a}(x)$. In this case, the factor marginals are given by

$$\begin{aligned} \mu_i(x) &\propto f_i(x) \prod_{b\in F(i)} \hat{\mu}^*_{b\to i}(x) \\ \mu_{V(a)}(\underline{x}_{V(a)}) &\propto f_a(\underline{x}_{V(a)}) \prod_{i\in V(a)} \mu^*_{i\to a}(x_i). \end{aligned} \tag{4}$$

Another consequence of the factor graph not having cycles is that the joint distribution μ can be written as a function of the factor marginals. This is especially convenient given that these marginals are easily computed with the SPA. The following lemma makes this precise.

Lemma 3.1. *Consider a factor graph without cycles. Let A be any subset of factor nodes whose induced subgraph is connected and let $V(A) \triangleq \cup_{a\in A} V(a)$ denote the set of variable nodes adjacent to A. Then, the marginal $\mu_{V(A)}$ can be written as*

$$\mu_{V(A)}(\underline{x}_{V(A)}) = \left(\prod_{a\in A} \frac{\mu_{V(a)}(\underline{x}_{V(a)})}{\prod_{i\in V(a)} \mu_i(x_i)}\right)\left(\prod_{i\in V(A)} \mu_i(x_i)\right).$$

Proof. The proof is by induction on $|A|$. If $|A| = 1$, then let b denote the single factor node in A and observe that the base case

$$\mu_{V(b)}(\underline{x}_{V(b)}) = \left(\frac{\mu_{V(b)}(\underline{x}_{V(b)})}{\prod_{i\in V(b)} \mu_i(x_i)}\right)\left(\prod_{i\in V(b)} \mu_i(x_i)\right) = \mu_{V(b)}(\underline{x}_{V(b)})$$

holds trivially. The subgraph, $S(A)$, induced by A is a tree because it is a connected subgraph of a cycle free graph. If $|A| > 1$, then choose $b \in A$ to be any factor node with $|V(b)| \geq 2$ that is adjacent to a leaf variable node. Such a b exists because $S(A)$ is a tree and $|A| > 1$. Since $S(A)$ is a tree, there is a unique variable node $k \in V(b)$ that is in both $V(b)$ and $V(A\backslash b)$. In this case, $S(V(A\backslash b))$ is connected and $|A\backslash b| = |A| - 1$. Therefore, we can apply the induction hypothesis to get the formula for $\mu_{V(A\backslash b)}$. Since x_k separates $V(A\backslash b)$ and $V(b)$ in the factor graph, conditional independence implies

$$\begin{aligned}
\mu_{V(A)}(\underline{x}_{V(A)}) &= \mu_{V(A\backslash b)}(\underline{x}_{V(A\backslash b)})\mu_{V(b)\backslash k}(\underline{x}_{V(b)\backslash k|\{k\}}|x_k) \\
&= \mu_{V(A\backslash b)}(\underline{x}_{V(A\backslash b)})\frac{\mu_{V(b)}(\underline{x}_{V(b)})}{\mu_k(x_k)} \\
&= \left(\prod_{a\in A\backslash b} \frac{\mu_{V(a)}(\underline{x}_{V(a)})}{\prod_{i\in V(a)} \mu_i(x_i)}\right)\left(\prod_{i\in V(A\backslash b)} \mu_i(x_i)\right)\frac{\mu_{V(b)}(\underline{x}_{V(b)})}{\mu_k(x_k)} \\
&= \left(\prod_{a\in A} \frac{\mu_{V(a)}(\underline{x}_{V(a)})}{\prod_{i\in V(a)} \mu_i(x_i)}\right)\left(\prod_{i\in V(b)} \mu_i(x_i)\right)\left(\prod_{i\in V(A\backslash b)} \mu_i(x_i)\right)\frac{1}{\mu_k(x_k)} \\
&= \left(\prod_{a\in A} \frac{\mu_{V(a)}(\underline{x}_{V(a)})}{\prod_{i\in V(a)} \mu_i(x_i)}\right)\left(\prod_{i\in V(A)} \mu_i(x_i)\right),
\end{aligned}$$

where the last equality holds because $V(A\backslash b) \cap V(b) = \{k\}$ implies that the second and third products have an extra factor of $\mu_k(x_k)$ that cancels the $1/\mu_k(x_k)$ term. □

Remark 3.2. Choosing $A = F$ in the above lemma shows that the formula holds for any tree factor graph. Likewise, it is easy to verify that conditional independence implies the formula also holds for any factor graph without cycles (i.e., consisting of disjoint tree components).

Lemma 3.3. *For a factor graph without cycles, the entropy of μ can be written in terms of the factor marginals as*

$$H(\mu) = \sum_{a\in F} H(\mu_{V(a)}) - \sum_{i\in V} (|F(i)| - 1)\, H(\mu_i) \tag{5}$$

and the free energy can be written as

$$-\ln Z = \sum_{a\in A}\left[\sum_{\underline{x}_{V(a)}} \mu_{V(a)}(\underline{x}_{V(a)}) \ln \frac{1}{f_a(\underline{x}_{V(a)})}\right] + \sum_{i\in V}\left[\sum_{x_i} \mu_i(x_i) \ln \frac{1}{f_i(x_i)}\right] - H(\mu). \tag{6}$$

Proof. Using the form of $\mu(\underline{x})$ given by Lemma 3.1, one can directly compute the entropy with

$$\begin{aligned}
H(\mu) &= \sum_{\underline{x}\in\mathcal{X}^n} \mu(\underline{x}) \ln \frac{1}{\mu(\underline{x})} \\
&= -\sum_{\underline{x}\in\mathcal{X}^n} \mu(\underline{x}) \ln \left[\left(\prod_{a\in F} \frac{\mu_{V(a)}(\underline{x}_{V(a)})}{\prod_{i\in V(a)} \mu_i(x_i)} \right) \left(\prod_{i\in V} \mu_i(x_i) \right) \right] \\
&= -\sum_{\underline{x}\in\mathcal{X}^n} \mu(\underline{x}) \left[\sum_{a\in F} \ln \mu_{V(a)}(\underline{x}_{V(a)}) - \sum_{i\in V} (|F(i)|-1) \ln \mu_i(x_i) \right] \\
&= \sum_{a\in F} \sum_{\underline{x}\in\mathcal{X}^n} \mu(\underline{x}) \ln \frac{1}{\mu_{V(a)}(\underline{x}_{V(a)})} - \sum_{i\in V} (|F(i)|-1) \sum_{\underline{x}\in\mathcal{X}^n} \mu(\underline{x}) \ln \frac{1}{\mu_i(x_i)} \\
&= \sum_{a\in F} H(\mu_{V(a)}) - \sum_{i\in V} (|F(i)|-1)\, H(\mu_i).
\end{aligned}$$

The second result follows by combining the result $G(\mu) = -\ln Z$ from (2) with (1). □

3.2 Factor Graphs with Cycles

For factor graphs with cycles, there is no guarantee that the SPA will converge or give useful results. Still, one can use equations that are exact for factor graphs without cycles as approximations for arbitrary factor graphs and hope for the best. This approach is sometimes called the *Bethe formalism*. Roughly speaking, the idea is to identify fixed points of the SPA and, at each fixed point, use (4) and Lemma 3.1 to estimate $\mu(\underline{x})$. Under various conditions, this can give good results. For example, if the SPA has a unique fixed point and the factor graph has large girth, then the marginal of any node is essentially determined by its tree-like neighborhood on which the SPA is exact.

Consider a set of variable node beliefs $b \triangleq \{b_i : \mathcal{X} \to [0,1]\}_{i\in V}$ and factor node beliefs $\hat{b} \triangleq \left\{\hat{b}_a : \mathcal{X}^{|V(a)|} \to [0,1]\right\}_{a\in F}$ satisfying the marginal consistency constraints

$$\sum_{\underline{x}_{V(a)\setminus i}} \hat{b}_a(\underline{x}_{V(a)}) = b_i(x_i),$$

for all $(i,a) \in E$ and $x_i \in \mathcal{X}$. The set of all $(b,\hat{b})$ pairs satisfying these conditions is called the *marginal polytope* associated with the factor graph and denoted by $\mathcal{M}$. A set of factor node beliefs $\hat{b}$ is called *consistent* if there is a b such that $(b,\hat{b}) \in \mathcal{M}$ and the set of all such beliefs is denoted by $\mathcal{M}'$. For any $\hat{b} \in \mathcal{M}$, the *Bethe entropy* is defined to be

$$H_B(\hat{b}) \triangleq \sum_{a\in F} \sum_{\underline{x}_{V(a)}} \hat{b}_a(\underline{x}_{V(a)}) \ln \frac{1}{\hat{b}_a(\underline{x}_{V(a)})} - \sum_{i\in V} (|F(i)|-1) \sum_{x_i} b_i(x_i) \ln \frac{1}{b_i(x_i)},$$

where we note that the b_i's can be computed from the $\hat{b}_a$'s. Basically, this formula treat the beliefs as marginals and applies the entropy formula (5) for a factor graph without cycles. This extends the domain of the entropy formula in (5) from factor graphs without cycles to general factor graphs.

Likewise, one can extend the free energy formula in (6) from factor graphs without cycles to general factor graphs. This results in the *Bethe free energy* formula

$$F_B(\hat{b}) = \sum_{a\in F} \sum_{\underline{x}_{V(a)}} \hat{b}_a(\underline{x}_{V(a)}) \ln \frac{1}{f_a(\underline{x}_{V(a)})} + \sum_{i\in V} \sum_{x_i} b_i(x_i) \ln \frac{1}{f_i(x_i)} - H_B(\hat{b}). \tag{7}$$

Reasoning by analogy from the variational Gibbs free energy, one might hope that minimizing this function over all consistent beliefs will lead to a set of beliefs that approximates well the marginals $\mu_{V(a)}$. Hence, we define

$$\hat{b}^* = \arg\min_{\hat{b}\in\mathcal{M}'} F_B(\hat{b})$$

and note that $F_B(\hat{b}^*)$ is known as the Bethe estimate of $-\ln Z$ because $\min_{\nu\in\mathcal{P}(\mathcal{X}^n)} G(\nu) = -\ln Z$. This leads us to the question, "How hard is minimizing the Bethe free energy?". In the next section, we will see that this question is intimately connected to the SPA.

Remark 3.4. The negation of the Bethe free energy is called the *Bethe free entropy* and can be similarly maximized. An easy way to remember the difference is to recall that physical processes tend to minimize energy and maximize entropy.

4 The Bethe Free Entropy and the SPA

In this section, we describe the dynamic formulation (9) of the Bethe free entropy in terms of the sum-product messages. If the messages are a fixed point of the SPA, then the dynamic formulation equals the Bethe free entropy associated with the sum-product beliefs given by (4). Otherwise, the function has no direct connection to the Bethe free energy. Instead, it defines the dynamics of the SPA because the SPA updates are defined naturally in terms of its derivatives.

4.1 The Bethe Free Energy at a SPA Fixed Point

Let the message vectors $\mu, \hat{\mu}$ define a fixed point of the SPA and, based on (4), let us define the node and variable beliefs to be

$$b_i(x_i) \triangleq \frac{1}{C_i} f_i(x_i) \prod_{a\in F(i)} \hat{\mu}_{a\to i}(x_i)$$
$$\hat{b}_a(\underline{x}_{V(a)}) \triangleq \frac{1}{\hat{C}_a} f_a(\underline{x}_{V(a)}) \prod_{i\in V(a)} \mu_{i\to a}(x_i),$$

where $C_i \triangleq \sum_{x_i} f_i(x_i) \prod_{a\in F(i)} \hat{\mu}_{a\to i}(x_i)$ and $\hat{C}_a \triangleq \sum_{x_{V(a)}} f_a(\underline{x}_{V(a)}) \prod_{i\in V(a)} \mu_{i\to a}(x_i)$. Also, for any $a \in F(i)$, one can observe that

$$\begin{aligned} f_i(x_i) \prod_{b\in F(i)} \hat{\mu}_{b\to i}(x_i) &= \hat{\mu}_{a\to i}(x_i) f_i(x_i) \prod_{b\in F(i)\setminus a} \hat{\mu}_{b\to i}(x_i) \\ &\propto \hat{\mu}_{a\to i}(x_i)\mu_{i\to a}(x_i). \end{aligned}$$

Thus, for $a \in F(i)$, one has

$$b_i(x_i) = \frac{1}{C_{i,a}} \hat{\mu}_{a\to i}(x_i)\mu_{i\to a}(x_i),$$

where $C_{i,a} \triangleq \sum_{x_i} \hat{\mu}_{a\to i}(x_i)\mu_{i\to a}(x_i)$. Using this, one can verify the marginal consistency condition by observing that

$$\begin{aligned} \sum_{\underline{x}_{V(a)\setminus j}} \hat{b}_a(\underline{x}_{V(a)}) &\propto \sum_{\underline{x}_{V(a)\setminus j}} f_a(\underline{x}_{V(a)}) \prod_{i\in V(a)} \mu_{i\to a}(x_i) \\ &\propto \mu_{j\to a}(x_j) \sum_{\underline{x}_{V(a)\setminus j}} f_a(\underline{x}_{V(a)}) \prod_{i\in V(a)} \mu_{i\to a}(x_i) \\ &\propto \mu_{j\to a}(x_j)\hat{\mu}_{a\to j}(x_j) \\ &\propto b_j(x_j). \end{aligned}$$

Now, we will write the Bethe free energy in terms of SPA fixed-point messages.. First, we expand the Bethe entropy term in (7) to get

$$F_B(\hat{b}) = -\sum_{a\in F} \sum_{\underline{x}_{V(a)}} \hat{b}_a(\underline{x}_{V(a)}) \ln \frac{f_a(\underline{x}_{V(a)})}{\hat{b}_a(\underline{x}_{V(a)})} - \sum_{i\in V} \sum_{x_i} b_i(x_i) \ln \frac{f_i(x_i)}{b_i(x_i)} + \sum_{i\in V} |F(i)| \sum_{x_i} b_i(x_i) \ln \frac{1}{b_i(x_i)}.$$

Using SPA fixed-point messages, the first term T_1 can be rewritten as

$$\begin{aligned} T_1 &\triangleq -\sum_{a\in F}\sum_{\underline{x}_{V(a)}} \hat{b}_a(\underline{x}_{V(a)}) \ln \frac{\hat{C}_a f_a(\underline{x}_{V(a)})}{f_a(\underline{x}_{V(a)}) \prod_{i\in V(a)} \mu_{i\to a}(x_i)} \\ &= -\sum_{a\in F}\sum_{\underline{x}_{V(a)}} \hat{b}_a(\underline{x}_{V(a)}) \left[\ln \hat{C}_a - \sum_{i\in V(a)} \ln \mu_{i\to a}(x_i)\right] \\ &= -\sum_{a\in F}\left[\ln \hat{C}_a - \sum_{i\in V(a)}\sum_{x_i} b_i(x_i) \ln \mu_{i\to a}(x_i)\right]. \end{aligned}$$

Using SPA fixed-point messages, the second term can be rewritten as

$$\begin{aligned} T_2 &\triangleq -\sum_{i\in V}\sum_{x_i} b_i(x_i) \ln \frac{C_i f_i(x_i)}{f_i(x_i) \prod_{a\in F(i)} \hat{\mu}_{a\to i}(x_i)} \\ &= -\sum_{i\in V}\sum_{x_i} b_i(x_i) \left[\ln C_i - \sum_{a\in F(i)} \ln \hat{\mu}_{a\to i}(x_i)\right] \\ &= -\sum_{i\in V}\left[\ln C_i - \sum_{a\in F(i)}\sum_{x_i} b_i(x_i) \ln \hat{\mu}_{a\to i}(x_i)\right]. \end{aligned}$$

Using SPA fixed-point messages, the third term can be rewritten as

$$\begin{aligned} T_3 &\triangleq \sum_{i\in V}\sum_{a\in F(i)}\sum_{x_i} b_i(x_i) \ln \frac{C_{i,a}}{\hat{\mu}_{a\to i}(x_i)\mu_{i\to a}(x_i)} \\ &= \sum_{i\in V}\sum_{a\in F(i)}\sum_{x_i} b_i(x_i) \left[\ln C_{i,a} - \ln \hat{\mu}_{a\to i}(x_i) - \ln \mu_{i\to a}(x_i)\right] \\ &= \sum_{i\in V}\sum_{a\in F(i)}\left[\ln C_{i,a} - \sum_{x_i} b_i(x_i) \ln \hat{\mu}_{a\to i}(x_i) - \sum_{x_i} b_i(x_i) \ln \mu_{i\to a}(x_i)\right]. \end{aligned}$$

Putting these three rewritten terms together, one gets

$$F_B(\hat{b}) = -\sum_{a\in F} \ln \hat{C}_a - \sum_{i\in V} \ln C_i + \sum_{(i,a)\in E} \ln C_{i,a}. \tag{8}$$

4.2 Dynamic Formulation of the Bethe Free Entropy

In this section, we observe that the RHS of (8) can be evaluated for any set of SPA messages (i.e., a fixed-point is not required). Of course, the previous derivation provides no meaning for such an evaluation. Fortunately, the SPA itself will provide the connection. One caveat is that there are multiple ways to write the Bethe free energy in terms of the SPA fixed-point messages and we have picked precisely the one for which this works. Note: To match the formulas in [2, Sec. 14.2.4], we now flip signs and discuss the Bethe free entropy $-F_B(\hat{b})$.

To express the dynamic formulation, one groups the messages by factor node and variable node and we use the notation $\mu_{\to a} \triangleq \{\mu_{i\to a}\}_{i\in V(a)}$ and $\hat{\mu}_{\to i} \triangleq \{\hat{\mu}_{a\to i}\}_{a\in F(i)}$. The potential functions associated

with the factor nodes, variable nodes, and edges are defined, respectively, by

$$S_a(\mu_{\to a}) \triangleq \ln\left[\sum_{\underline{x}_{V(a)}} f_a(\underline{x}_{V(a)}) \prod_{i\in V(a)} \mu_{i\to a}(x_i)\right]$$

$$\hat{S}_i(\hat{\mu}_{\to i}) \triangleq \ln\left[\sum_{x_i} f_i(x_i) \prod_{b\in F(i)} \hat{\mu}_{b\to i}(x_i)\right]$$

$$S_{i,a}(\mu_{i\to a}, \hat{\mu}_{a\to i}) \triangleq \ln\left[\sum_{x_i} \mu_{i\to a}(x_i)\hat{\mu}_{a\to i}(x_i)\right].$$

The *Bethe free entropy* for a set of SPA messages is defined to be the sum of these potentials over all nodes and edges

$$S(\mu, \hat{\mu}) \triangleq \sum_{a\in F} S_a(\mu_{\to a}) + \sum_{i\in V} \hat{S}_i(\hat{\mu}_{\to i}) - \sum_{(i,a)\in E} S_{i,a}(\mu_{i\to a}, \hat{\mu}_{a\to i}), \tag{9}$$

where $\mu \triangleq \{\mu_{i\to a}\}_{(i,a)\in E}$ and $\hat{\mu} \triangleq \{\hat{\mu}_{a\to i}\}_{(i,a)\in E}$. This definition is taken from [2, Sec. 14.2.4] though our notation differs slightly from theirs.

The following derivation of the connection between the Bethe free entropy and the sum-product algorithm is very much related to the Lagrangian approach of Yedidia et al. [3], but seems to avoid the fixed-point requirement by focusing on edge messages. We note that [3] proves, under some conditions, beliefs minimizing the static Bethe free energy $F_B(\hat{b})$ (or maximizing the static Bethe free entropy) are given by fixed points of the SPA. For details, see [2, Sec. 14.4.1].

First, we note that $S(\mu, \hat{\mu})$ is unaffected by the positive scaling of individual messages. This is because scaling any message $\mu_{i\to a}(x)$ (resp. $\hat{\mu}_{a\to i}(x)$) by a constant $\gamma > 0$ adds $\ln\gamma$ to $S_a(\mu_{\to a})$ (resp. $\hat{S}_i(\hat{\mu}_{\to i})$) and adds $\ln\gamma$ to $S_{i,a}(\mu_{i\to a}, \hat{\mu}_{a\to i})$. It is easy to verify that these two constants cancel in (9) and thus $S(\mu, \hat{\mu})$ is unchanged.

Now, consider the derivative of $S(\mu, \hat{\mu})$ w.r.t. $\mu_{j\to a}(x)$. The individual terms are given by

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\mu_{j\to a}(x)} \sum_{b\in F} S_b(\mu_{\to b}) &= \frac{\mathrm{d}}{\mathrm{d}\mu_{j\to a}(x)} S_a(\mu_{\to a}) \\
&= \frac{\frac{\mathrm{d}}{\mathrm{d}\mu_{j\to a}(x)} \sum_{\underline{x}_{V(a)}} f_a(\underline{x}_{V(a)}) \prod_{i\in V(a)} \mu_{i\to a}(x_i)}{\sum_{\underline{x}_{V(a)}} f_a(\underline{x}_{V(a)}) \prod_{i\in V(a)} \mu_{i\to a}(x_i)} \\
&= \frac{\sum_{\underline{x}_{V(a)}} f_a(\underline{x}_{V(a)})\delta_{x_j,x} \prod_{i\in V(a)\setminus j} \mu_{i\to a}(x_i)}{\sum_{x_j} \mu_{j\to a}(x_j) \sum_{\underline{x}_{V(a)\setminus j}} f_a(\underline{x}_{V(a)}) \prod_{i\in V(a)\setminus j} \mu_{i\to a}(x_i)}
\end{aligned}$$

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\mu_{j\to a}(x)} \sum_{(i,b)\in E} S_{i,b}(\mu_{i\to b}, \hat{\mu}_{b\to i}) &= \frac{\mathrm{d}}{\mathrm{d}\mu_{j\to a}(x)} S_{j,a}(\mu_{j\to a}, \hat{\mu}_{a\to j}) \\
&= \frac{\frac{\mathrm{d}}{\mathrm{d}\mu_{j\to a}(x)} \sum_{x_j} \mu_{j\to a}(x_j)\hat{\mu}_{a\to j}(x_j)}{\sum_{x'} \mu_{j\to a}(x')\hat{\mu}_{a\to j}(x')} \\
&= \frac{\hat{\mu}_{a\to j}(x)}{\sum_{x'} \mu_{j\to a}(x')\hat{\mu}_{a\to j}(x')}.
\end{aligned}$$

In this case, one finds that

$$\frac{\mathrm{d}}{\mathrm{d}\mu_{j\to a}(x)} S(\mu, \hat{\mu}) = \frac{\sum_{\underline{x}_{V(a)}} f_a(\underline{x}_{V(a)})\delta_{x_j,x} \prod_{i\in V(a)\setminus j} \mu_{i\to a}(x_i)}{\sum_{x_j} \mu_{j\to a}(x_j) \sum_{\underline{x}_{V(a)\setminus j}} f_a(\underline{x}_{V(a)}) \prod_{i\in V(a)\setminus j} \mu_{i\to a}(x_i)} - \frac{\hat{\mu}_{a\to j}(x)}{\sum_{x'} \mu_{j\to a}(x')\hat{\mu}_{a\to j}(x')}. \tag{10}$$

Setting this derivative to 0 and solving for $\hat{\mu}_{a\to j}(x)$ is equivalent to solving for $\underline{c}$ in

$$\frac{b_j}{\sum_i a_i b_i} - \frac{c_j}{\sum_i a_i c_i} = 0,$$

when $\underline{a}$ and $\underline{b}$ are fixed. It is not too hard to see that this holds iff c_j/b_j is constant for all j. Therefore, as a function of $\hat{\mu}_{a\to j}(x)$, (10) is zero for all $x \in \mathcal{X}$ if and only if

$$\hat{\mu}_{a\to j}(x) \propto \sum_{\underline{x}_{V(a)}} f_a(\underline{x}_{V(a)})\delta_{x_j,x} \prod_{i\in V(a)\setminus j} \mu_{i\to a}(x_i).$$

This condition is precisely equal to the sum-product update for $\hat{\mu}_{a\to j}(x)$.

Likewise, the derivative of $S(\mu,\hat{\mu})$ w.r.t. $\hat{\mu}_{a\to j}(x)$ is related to the sum-product update. The individual terms are given by

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\hat{\mu}_{a\to j}(x)} \sum_{i\in V} \hat{S}_i(\hat{\mu}_{\to i}) &= \frac{\mathrm{d}}{\mathrm{d}\hat{\mu}_{a\to j}(x)} \hat{S}_j(\hat{\mu}_{\to j}) \\
&= \frac{\frac{\mathrm{d}}{\mathrm{d}\hat{\mu}_{a\to j}(x)} \sum_{x_j} f_j(x_j) \prod_{b\in F(j)} \hat{\mu}_{b\to j}(x_j)}{\sum_{x_j} f_j(x_j) \prod_{b\in F(j)} \mu_{j\leftarrow b}(x_j)} \\
&= \frac{\sum_{x_j} f_j(x_j) \prod_{b\in F(j)\setminus a} \hat{\mu}_{b\to j}(x_j)\delta_{x_j,x}}{\sum_{x_j} f_j(x_j) \prod_{b\in F(j)} \hat{\mu}_{b\to j}(x_j)} \\
&= \frac{f_j(x) \prod_{b\in F(j)\setminus a} \hat{\mu}_{b\to j}(x)}{\sum_{x_j} \hat{\mu}_{a\to j}(x_j) f_j(x_j) \prod_{b\in F(j)\setminus a} \hat{\mu}_{b\to j}(x_j)}
\end{aligned}$$

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\hat{\mu}_{a\to j}(x)} \sum_{(i,b)\in E} S_{i,b}(\mu_{i\to b},\mu_{i\leftarrow b}) &= \frac{\mathrm{d}}{\mathrm{d}\hat{\mu}_{a\to j}(x)} S_{j,a}(\mu_{j\to a},\hat{\mu}_{a\to j}) \\
&= \frac{\frac{\mathrm{d}}{\mathrm{d}\hat{\mu}_{a\to j}(x)} \sum_{x_j} \mu_{j\to a}(x_j)\hat{\mu}_{a\to j}(x_j)}{\sum_{x_j} \mu_{j\to a}(x_j)\hat{\mu}_{a\to j}(x_j)} \\
&= \frac{\mu_{j\to a}(x)}{\sum_{x_j} \mu_{j\to a}(x_j)\hat{\mu}_{a\to j}(x_j)}
\end{aligned}$$

This implies that

$$\frac{\mathrm{d}}{\mathrm{d}\hat{\mu}_{a\to j}(x)} S(\mu,\hat{\mu}) = \frac{f_j(x) \prod_{b\in F(j)\setminus a} \hat{\mu}_{b\to j}(x)}{\sum_{x_j} \hat{\mu}_{a\to j}(x_j) f_j(x_j) \prod_{b\in F(j)\setminus a} \hat{\mu}_{b\to j}(x_j)} - \frac{\mu_{j\to a}(x)}{\sum_{x_j} \mu_{j\to a}(x_j)\hat{\mu}_{a\to j}(x_j)}. \tag{11}$$

Using the same argument as before, one finds that (11), as a function of $\mu_{j\to a}(x)$, is zero for all $x \in \mathcal{X}$ if and only if

$$\mu_{j\to a}(x) \propto f_j(x) \prod_{b\in F(j)\setminus a} \hat{\mu}_{b\to j}(x).$$

This condition is precisely equal to the sum-product update for $\mu_{j\to a}(x)$.

Thus, the SPA alternately updates the forward and backward messages in a way that forces the gradient (w.r.t. the opposite set of messages) to the all zero vector. For this reason, we interpret (9) as a function that generates the discrete-time dynamics of the SPA.

References

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley Series in Telecommunications, Wiley, 2nd. ed., 2006.

[2] M. Mezard and A. Montanari, *Information, Physics, and Computation*. New York, NY: Oxford University Press, 2009.

[3] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Constructing free energy approximations and generalized belief propagation algorithms," *IEEE Trans. Inform. Theory*, vol. 51, pp. 2282–2312, 2005.