Stat 375 Inference in Graphical Models

Variational Methods: A Short Overview

Andrea Montanari *Lecture 6-7 - 2/2/2009*

The following synthesis is based on notes taken in 2007 by Fernando Amat, Argyris Zymnis, John Duchi, Krishnamurthy Iyer, Vahideh H. Manshadi and Sewoong Oh. The merging might have lead to inconsistencies that I will repair in the version posted online. Excellent (and more complete) references on variational methods are the review [WJ08], as well as the papers [WJW05] and [YFW05]. Relevant material can also be found in Chapters 4 and of [MM09]

1 Free Energy and Gibbs Free Energy

The basic idea in variational methods is to reduce the problem of counting/marginalizing a distribution to an optimization problem. The function to be optimized is called free energy (the name comes from statistical physics). We start by defining free energy and the Gibbs free energy of a graphical model. As before $G = (V = [n], F, E)$ will be a factor graph, $\{\psi_a(\,\cdot\,)\}_{a\in F}$ the compatibility functions. We further let $\psi(\underline{x}) \equiv \prod_{a\in F}\psi_a(\underline{x}_{\partial a})$.

Definition 1 ((Helmotz) Free Energy)**.** *Given a factor graph model on $\underline{x}$*

$$\mu(\underline{x}) = \frac{1}{Z}\prod_{a\in\mathcal{F}}\psi_a(x_{\partial a}),$$

its (Helmotz) free energy is defined as

$$\Phi \equiv \log Z.$$

We already saw that counting is equivalent to computing Φ and that computing marginals of the variables is equivalent to computing differences $\Phi - \Phi'$.

Definition 2 (Gibbs free energy)**.** *Given a distribution $b \in \mathsf{M}(\mathcal{X}^V)$, its Gibbs free energy is defined as*

$$\begin{aligned}\mathbb{G}[b] &\equiv H[b] + \mathbb{E}_b \log\psi(\underline{x}) &(1)\\ &= -\sum_{\underline{x}} b(\underline{x})\log b(\underline{x}) + \sum_{\underline{x}} b(\underline{x})\log\psi(\underline{x})\,. &(2)\end{aligned}$$

Proposition 3. *The function $\mathbb{G} : \mathsf{M}(\mathcal{X}^V) \to \mathbb{R}$ is strictly concave on the convex domain $\mathsf{M}(\mathcal{X}^V)$. Further its unique maximum is achieved at $b = \mu$, with $\mathbb{G}[\mu] = \Phi$.*

Proof Concavity is just a consequence of the fact that $z \mapsto z\log z$ is convex on $\mathbb{R}_+$. The unique minimum can be found by introducing the Lagrangian

$$\mathcal{L}(b,\lambda) = \mathbb{G}[b] - \lambda\Big(\sum_{\underline{x}} b(\underline{x}) - 1\Big)\,. \tag{3}$$

Setting to zero the derivative with respect to $b(\underline{x})$, one gets $b(\underline{x}) = \psi(\underline{x})/Z$. It is immediate to check that the value at the minimum is indeed Φ. □

Observation 4. *Gibbs free energy can be re-expressed as follows:*

$$G[b] = \Phi - D(b||\mu)$$

where $D(n||\mu)$ is the Kullback-Leibler divergence between b and μ (from information theory).

We reduced marginal computation to a convex optimization problem. The problem is that the search space is too big. Indeed $\mathsf{M}(\mathcal{X}^V)$ is a simplex of dimensions $(|\mathcal{X}|^n - 1)$.

2 Naive mean field

The simplest approach to overcome the dimensionality problem is to restrict the measure $b(\underline{x})$ to belong to a simple family or subset of $\mathsf{M}(\mathcal{X}^V)$. The problem, as we will see, is that some subsets of $M(\mathcal{X}^V)$ might not be convex, so we lose the nice property of having a convex problem.

The simplest approximation scheme consists in taking b to be a factorized distribution: $b(\underline{x}) = \prod_{i\in V} b_i(x_i)$. The resulting submanifold of $\mathsf{M}(\mathcal{X}^V)$ has dimension $n(|\mathcal{X}|-1)$ but is non-convex. Plugging this factorized form in the Gibbs free energy expression, we obtain the naive mean-field Gibbs free energy:

$$\begin{aligned}\mathbb{G}_{\mathrm{MF}}(b) &= H[b] + \mathbb{E}_\nu[\log\psi(\underline{x})] \\ &= -\sum_{i\in V}\sum_{x_i} b_i(x_i)\log b_i + \sum_{a\in F}\sum_{\underline{x}_{\partial a}}\prod_{i\in\partial a} b_i(x_i)\log\psi_a(\underline{x}_{\partial a})\end{aligned}$$

We want now to optimize this over b_i. We can use Lagrange multipliers λ_i to constrain the mean field marginals to be normalized, that is, $\sum_{x_i} b_i(x_i) = 1$, giving us the following Lagrangian and derivatives:

$$\begin{aligned}\mathcal{L}(\underline{b},\lambda) &= -\sum_{i\in V}\sum_{x_i} b_i(x_i)\log b_i(x_i) + \sum_{a\in F}\sum_{\underline{x}_{\partial a}}\prod_{i\in\partial a} b_i(x_i)\log\psi_a(\underline{x}_{\partial a}) + \sum_{i\in V}\lambda_i\left(\sum_{x_i} b_i(x_i) - 1\right) \\ \frac{\partial\mathcal{L}(\underline{b},\lambda)}{\partial b_i(x_i)} &= -1 - \log b_i(x_i) + \sum_{a\in F}\sum_{x_j : j\in\partial a\setminus i}\prod_{j\in\partial a\setminus i} b_j(x_j)\log\psi_a(\underline{x}_{\partial a}) + \lambda_i\end{aligned}$$

Solving, we see that stationary points must satisfy

$$\begin{aligned}b_i(x_i) &= \exp\left(\sum_{a\in F}\sum_{x_j : j\in\partial a\setminus i}\prod_{j\in\partial a\setminus i} b_j(x_j)\log\psi_a(\underline{x}_{\partial a}) + \lambda_i - 1\right) \\ &= \frac{1}{Z_i}\exp\left(\sum_{a\in F}\sum_{x_j : j\in\partial a\setminus i}\prod_{j\in\partial a\setminus i} b_j(x_j)\log\psi_a(\underline{x}_{\partial a})\right) \qquad (4)\end{aligned}$$

where we solved for λ_i to normalize the b's. This suggests an iterative algorithm for optimizing the b's by iterating the above equation until the b's converge.

2.1 Example: Ising Models

As an example, we will be considering ferromagnetic Ising models, which can be thought of as mathematical models of the alignment of localized magnetic moments in a ferromagnetic material.

Given a pairwise graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, variables x_i's are associated to its vertices $i \in \mathcal{V}$. In an Ising model $x_i \in \mathcal{X} = \{+1, -1\}$. We have a factor over every "adjacent" (meaning there is only one factor node between the variable nodes) pair of variables as follows:

$$\psi_{ij}(x_i, x_j) = e^{\beta x_i x_j}, \quad \text{i.e.} \quad \psi_{ij} = \begin{pmatrix} e^{\beta} & e^{-\beta} \\ e^{-\beta} & e^{\beta} \end{pmatrix}.$$

This defines a distribution

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

To use the mean field approximation on an Ising model, we parametrize singleton marginals using their means $\mathbb{E}_{b_i}[x_i] \equiv m_i$:

$$b_i(x_i) = \frac{1 + m_i x_i}{2}.$$

The naive mean field free energy is then

$$\begin{aligned}\mathbb{G}_{\mathrm{MF}}(\underline{m}) &= \sum_{i\in V} H(b_i) + \beta \sum_{(i,j)\in\mathcal{E}} \sum_{x_i x_j} b_i(x_i) b_j(x_j) x_i x_j \\ &= \sum_{i\in V} h(m_i) + \beta \sum_{(i,j)\in\mathcal{E}} m_i m_j,\end{aligned}$$

where $h(m) \equiv -\frac{1+m}{2}\log(\frac{1+m}{2}) - \frac{1-m}{2}\log(\frac{1-m}{2})$.

Using equation (4) for mean fields, we find that

$$b_i(x_i) = \frac{1+m_i x_i}{2} \propto \exp\Big(\sum_{j\in\partial i}\sum_{x_j} b_j(x_j)\beta x_i x_j\Big) = \exp\Big(\beta x_i \sum_{j\in\partial i} \mathbb{E}_{\nu_j}[x_j]\Big) = \exp\Big(\beta x_i \sum_{j\in\partial i} m_j\Big).$$

and we thus have the iterative updates

$$\begin{aligned} m_i &= \mathbb{E}_{b_i}[x_i] \\ &= \frac{\exp\left(\beta\sum_{j\in\partial i} m_j\right) - \exp\left(-\beta\sum_{j\in\partial i} m_j\right)}{\exp\left(\beta\sum_{j\in\partial i} m_j\right) + \exp\left(-\beta\sum_{j\in\partial i} m_j\right)} \\ &= \tanh\Big(\beta\sum_{j\in\partial i} m_j\Big). \end{aligned} \tag{5}$$

The last line is the hyperbolic tangent, which is a strictly increasing function bounded by $[-1, 1]$.

2.2 Convergence of mean-field updates for Ising models

In this section, we will be considering the convergence properties of the iterative updates we have derived for our mean-field approximation of Ising models. For the sake of simplicity, we assume that $\mathcal{G}$ is regular, with degree k.

First, we observe that – under the update rules in Eq. (5) – if we initialize $m_i^{(0)} = m^{(0)}$ for all $i \in V$, we will have $m_i^{(t)} = m^{(t)}$ for all $t \geq 0$ (here $m_i^{(t)}$ is the value of m_i at the t^{th} iterative update). In other words, homogeneous estimates remain homogeneous under the update. The marginals evolve as follow

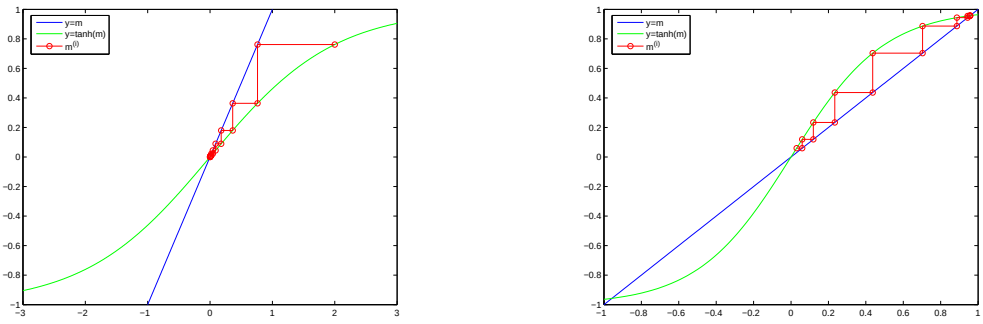
$$m^{(t+1)} = \tanh(\beta k\, m^{(t)}) \equiv f(m^{(t)})\,. \tag{6}$$

If $\beta < \frac{1}{k}$, then the function f is Lipschitz with constant smaller than one, so that no matter the value of $m^{(0)}$, the iterations will converge to $\lim_{t\to\infty} m^{(t)} = 0$. An illustration of this is in Fig. 1(a). If $\beta > \frac{1}{k}$, on the other hand, the slope of f will be greater than 1 at $m = 0$, so that there will be three points that the line $y = m$ will intersect $y = \tanh(\beta k m)$. Thus, if we begin the iterations with $m^{(0)} > 0$, $\lim_{t\to\infty} m_i^{(t)} = +m^*$. If $m^{(0)} < 0$, then the iterations will converge to $-m^*$, and in the degenerate case that $m^{(0)} = 0$, we will converge to the saddlepoint 0. An illustartion of this is in Fig. 1(b).

2.3 Marginals in the Ising model

Up ot this point, we have ignored the marginals of the actual distribution of the Ising model, focusing instead on the partition function and maximizing the free energy. What can we say about the marginals of the distribution that we obtain from our mean field approach?

It is relatively easy to see that the real marginals should be $\mu_i(+1) = \mu_i(-1) = \frac{1}{2}$, because the factors are completely symmetric. Nevertheless, the presence of two (positive and negative) attractive fixed points is a trace of an interesting phenomenon. For many graphs $\mathcal{G}$, at large enough β, the model undergoes a *phase transition*. More precisely, a typical configuration $\underline{x}$ distributed according $\mu(\,\cdot\,)$ is polarized: either $x_i = +1$



(a) (b)

Figure 1: (a) Updates to $m^{(t)}$ with $\beta < \frac{1}{2d}$. Path is down and left. (b) Updates to $m^{(t)}$ with $\beta > \frac{1}{2d}$. Path is up and right.

for about $n(1+m_+)/2$ vertices i, or $x_i = -1$ for about $n(1+m_+)/2$ vertices i, with $m_+ = m_+(\beta)$ strictly positive. Each of the two events occurr with the same probability. Further, marginals over several variable node that are far apart have a distribution that approximately decomposes as follows:

$$\mu(x_i, x_j, x_k) \approx \frac{1}{2}(\mu^+(x_i)\mu^+(x_j)\mu^+(x_k) + \mu^-(x_i)\mu^-(x_j)\mu^-(x_k)),$$

where $\mu^+(x_i)$ has positive mean, and $\mu^-(x_i)$ negative mean. This gives intuition as to why there are two distinct fixed points for strongly correlated factors–almost all of the nodes should be $+1$ or -1.

3 Bethe Free Energy

Is belief propagation associated to any variational principle? The answer is yes, although the relation is not as straightforward as for naive mean field. The associated free energy can be written in two forms that, as we will see, are related by a lagrangian duality: the first form is in terms of messages $\{\nu_{i\to a}(\,\cdot\,), \widehat{\nu}_{a\to i}(\,\cdot\,)\}$; the second in terms of marginals $\{b_i(\,\cdot\,), b_a(\,\cdot\,)\}$. We shall refer to both forms as to the *Bethe free energy*.

Let us start by defining the Bethe free energy $\mathbb{G}^*_{\mathrm{B}} : \{\nu, \hat{\nu}\} \to \mathbb{R}$ in terms of messages:

$$\begin{aligned}\mathbb{G}^*_{\mathrm{B}}(\underline{\nu}) \;=\; & -\sum_{i,a} \log\left(\sum_{x_i} \nu_{i\to a}(x_i)\hat{\nu}_{a\to i}(x_i)\right) + \sum_{a\in F} \log\left(\sum_{\underline{x}_{\partial a}} \psi_a(\underline{x}_{\partial a}) \prod_{i\in\partial a} \nu_{i\to a}(x_i)\right) \\ & + \sum_{i\in V} \log\left(\sum_{x_i} \prod_{a\in\partial i} \hat{\nu}_{a\to i}(x_i)\right).\end{aligned}$$

Note that there are three terms in the expression, one for each edge, one for each factor node, and one for each variable node in the factor graph. The following claim justifies the study of Bethe free energy.

Proposition 5. *Consider a factor graph model with strictly positive compatibility functions $\psi_a(\underline{x}_{\partial a})$. Then stationary points of the Bethe free energy are fixed points of the Belief Propagation algorithm and vice versa. That is,*

$$\frac{\partial \mathbb{G}^*_{\mathrm{B}}}{\partial \underline{\nu}} = 0$$

if and only if the messages are a fixed point of belief propagation.

Proof Note first from the definition of the Bethe free energy, that it is independent of scale with respect to each of its arguments. That is, if we replace, $\nu_{i\to a}(x_i)$ by $\lambda\nu_{i\to a}(x_i)$ for some $\lambda > 0$, then the value of the Bethe free energy remains the same. As a consequence, it is not necessary to impose the normalization constraints through Lagrange multipliers. Differentiating the expression for the Bethe Free energy with respect to $\nu_{i\to a}(x_i)$, we get

$$\frac{\partial \mathbb{G}^*_{\mathrm{B}}(\underline{\nu})}{\partial \nu_{i\to a}(x_i)} = -\frac{\hat{\nu}_{a\to i}(x_i)}{\sum_{x_i}\nu_{i\to a}(x_i)\hat{\nu}_{a\to i}(x_i)} + \frac{\sum_{\underline{x}_{\partial a\setminus i}}\psi_a(\underline{x}_{\partial a})\prod_{j\in\partial a\setminus i}\nu_{j\to a}(x_j)}{\sum_{\underline{x}_{\partial a}}\psi_a(\underline{x}_{\partial a})\prod_{j\in\partial a}\nu_{j\to a}(x_j)}$$

Setting $\frac{\partial \mathbb{G}^*_{\mathrm{B}}(\underline{\nu})}{\partial \nu_{i\to a}(x_i)} = 0$ and rearranging the expression we see that,

$$\begin{aligned}\hat{\nu}_{a\to i}(x_i) &= \left(\frac{\sum_{x_i}\nu_{i\to a}(x_i)\hat{\nu}_{a\to i}(x_i)}{\sum_{\underline{x}_{\partial a}}\psi_a(\underline{x}_{\partial a})\prod_{j\in\partial a}\nu_{j\to a}(x_j)}\right)\sum_{\underline{x}_{\partial a\setminus i}}\psi_a(\underline{x}_{\partial a})\prod_{j\in\partial a\setminus i}\nu_{j\to a}(x_j)\\ &\cong \sum_{\underline{x}_{\partial a\setminus i}}\psi_a(\underline{x}_{\partial a})\prod_{j\in\partial a\setminus i}\nu_{j\to a}(x_j)\end{aligned}$$

which is nothing but the equation for the fixed point of the Belief propagation algorithm (for factor message). The corresponding equation for the variable message can be obtained similarly by differentiating the expression for Bethe Free energy with respect to $\hat{\nu}_{a\to i}(x_i)$ and setting $\frac{\partial \mathbb{G}^*_{\mathrm{B}}}{\partial \hat{\nu}_{a\to i}(x_i)} = 0$. Doing this we get,

$$\nu_{i\to a}(x_i) \cong \prod_{b\in\partial i\setminus a}\hat{\nu}_{b\to i}(x_i)$$

Thus the stationary points of the Bethe free energy are the fixed points for the Belief Propagation algorithm. The converse can be obtained by direct substitution. □

4 Relation with marginals

The Naive Mean Field theory assumes that the beliefs over the entire graph is the product of the beliefs on the individual variable nodes.

$$\mu(\underline{x}) \simeq \prod_{i\in V}\mu_i(x_i)$$

This assumes that the values at the variable nodes are independent. On the other hand, the variables $x_{\partial a}$ might be strongly correlated. For example, one can have $\psi_a(x_{\partial a}) = \mathbb{I}_{\{x_i+x_j+x_k=0\}}$, where each variable assumes value in $\{0,1\}$.

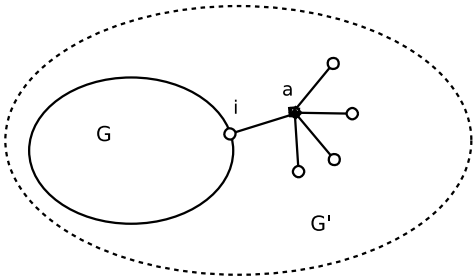


Figure 2: Addition of a single factor node to graph G

Lemma 6. *If G is a tree, then*

$$\begin{aligned}\mu(\underline{x}) &= \prod_{a\in F}\left(\frac{\mu_a(\underline{x}_{\partial a})}{\prod_{i\in\partial a}\mu_i(x_i)}\right)\prod_{i\in V}\mu_i(x_i)\\ &= \prod_{a\in F}\mu_a(\underline{x}_{\partial a})\prod_{i\in V}\mu_i(x_i)^{1-|\partial i|}\end{aligned}$$

Proof The proof will proceed by induction on $|F|$. To begin the induction step, we see that for $|F| = 0$, the given expression reduces to $\mu(\underline{x}) = \prod_{i\in V}\mu_i(x_i)$ which is nothing but the naive mean field equation.

Now assume that the given expression is true for $|F| \leq m$. We need to now check the expression for $|F| = m + 1$. Towards that end, assume we have a tree G with m factor nodes, and at variable node i, we append a factor node a (along with other variable nodes connected to a) at i, to get a tree G' with $m + 1$ factor nodes. See Fig(2) for details.

Note that $|\partial i|$ is the number of factor nodes connected to variable node i in graph G'. Thus number of factor nodes connected to i in graph G is $|\partial i| - 1$. By induction hypothesis,

$$\mu_(\underline{x}_G) = \prod_{b\in G}\mu_b(\underline{x}_{\partial b})\prod_{j\in G\setminus i}\mu_j(x_j)^{1-|\partial j|}\mu_i(x_i)^{1-(|\partial i|-1)}$$

Conditioning on $\underline{x}_G$, we get $\mu(\underline{x}_{G'}) = \mu(\underline{x}_G)\mu(\underline{x}_{\partial a\setminus i}|\underline{x}_G)$. By conditional independence (i.e. the Markov property), the above expression becomes,

$$\begin{aligned}\mu(\underline{x}_{G'}) &= \mu(\underline{x}_G)\mu(\underline{x}_{\partial a\setminus i}|x_i)\\ &= \prod_{b\in G}\mu_b(\underline{x}_{\partial b})\prod_{j\in G\setminus i}\mu_j(x_j)^{1-|\partial j|}\mu_i(x_i)^{1-(|\partial i|-1)}\frac{\mu(\underline{x}_{\partial a\setminus i},x_i)}{\mu_i(x_i)}\\ &= \prod_{b\in G'}\mu_b(\underline{x}_{\partial b})\prod_{j\in G'}\mu_j(x_j)^{1-|\partial j|}\end{aligned}$$

where the last equality follows from the fact that $\mu(\underline{x}_{\partial a \setminus i}, x_i) = \mu_a(\underline{x}_{\partial a})$ and for variable nodes $j \in \partial a \setminus i$ the number of neighbouring factor nodes $|\partial j| = 1$. This completes the induction step, thereby proving the lemma. □

As an example, consider a Markov chain. The expression for the distribution now reduces to

$$\begin{aligned}
\mu(\underline{x}) &= \prod_{a=1}^{n-1} \mu_a(x_a, x_{a+1}) \prod_{i=2}^{n-1} \mu_i(x_i)^{-1} \\
&= \mu_1(x_1, x_2) \prod_{i=2}^{n-1} \frac{\mu_i(x_i, x_{i+1})}{\mu_i(x_i)} \\
&= \mu_1(x_1)\mu_1(x_2|x_1) \prod_{i=2}^{n-1} \mu_i(x_{i+1}|x_i) \\
&= \mu_1(x_1) \prod_{i=1}^{n-1} \mu_i(x_{i+1}|x_i)
\end{aligned}$$

Thus we obtain the usual formula for the distribution of a Markov Chain.

Calculating the Gibbs energy for trees, one obtains,

$$\begin{aligned}
\mathbb{G}(\mu) &= H[\mu] + \mathbb{E}_\mu[\log(\psi(\underline{x}))] \\
&= -\sum_{\underline{x}} \mu(\underline{x}) \log(\mu(\underline{x})) + \sum_{a \in F} \sum_{\underline{x}_{\partial a}} \mu_a(\underline{x}_{\partial a}) \log(\psi_a(\underline{x}_{\partial a})) \\
&= -\sum_{\underline{x}} \mu(\underline{x}) \log \left(\prod_{a \in F} \mu_a(\underline{x}_{\partial a}) \prod_{i \in V} \mu_i(x_i)^{1-|\partial i|} \right) + \sum_{a \in F} \sum_{\underline{x}_{\partial a}} \mu_a(\underline{x}_{\partial a}) \log(\psi_a(\underline{x}_{\partial a})) \\
&= -\sum_{a \in F} \sum_{\underline{x}_{\partial a}} \mu_a(\underline{x}_{\partial a}) \log \mu_a(\underline{x}_{\partial a}) - \sum_{i \in V} \sum_{x_i} (1 - |\partial i|) \mu_i(x_i) \log(\mu_i(x_i)) + \sum_{a \in F} \sum_{\underline{x}_{\partial a}} \mu_a(\underline{x}_{\partial a}) \log(\psi_a(\underline{x}_{\partial a})) \\
&= \sum_{a \in F} H[\mu_a] - \sum_{i \in V} (|\partial i| - 1) H[\mu_i] + \sum_{a \in F} \sum_{\underline{x}_{\partial a}} \mu_a(\underline{x}_{\partial a}) \log(\psi_a(\underline{x}_{\partial a}))
\end{aligned}$$

Here $H[.]$ is the Shannon entropy.

5 Locally Consistent Marginals

For the general case, one can have marginals $\{b_i, b_a\}$ for each factor node $a \in F$ and variable node $i \in V$, which define the Gibbs free energy. We see that such marginals have to satisfy the following conditions to be a valid set of marginals.

$$\begin{aligned}
b_i(x_i) &\geq 0, \quad \forall x_i, \quad \forall i \in V \\
b_a(\underline{x}_{\partial a}) &\geq 0, \quad \forall \underline{x}_{\partial a}, \quad \forall a \in F \qquad (7) \\
\sum_{x_i} b_i(x_i) &= 1, \quad \forall i \in V \qquad (8) \\
\sum_{\underline{x}_{\partial a \setminus i}} b_a(\underline{x}_{\partial a}) &= b_i(x_i), \quad \forall a \in F, i \in \partial a \qquad (9) \\
\sum_{\underline{x}_{\partial a}} b_a(\underline{x}_{\partial a}) &= 1, \quad \forall a \in F. \qquad (10)
\end{aligned}$$

Note that Eqn(8) and Eqn(9) imply Eqn(10), and hence we can remove that equation from the requirements of a valid marginal.

The set of marginals $\{b_i, b_a\}$ on a graph G which satisfy the above requirements is known as the set of *locally consistent marginals* on G, denoted by $LOC(G)$. As the above constraints are linear, it can be readily seen that the set $LOC(G)$ is a convex set. Also the dimension of $LOC(G)$ is bounded above by $|\mathcal{X}|^{\max|\partial a|}|F|$.

We will now contrast the set of locally consistent marginals $LOC(G)$ for the cases when G is a tree and when G is not.

- G is a Tree
  - When G is a tree, then for any $\{b_i, b_a\} \in LOC(G)$, there exists an unique measure $b_* \in Measures(\mathcal{X}^\nu)$ whose marginals are given by $\{b_i, b_a\}$.
  - The measure b_* is given by

$$b_*(\underline{x}) = \prod_{a \in F} \left(\frac{b_a(\underline{x}_{\partial a})}{\prod_{i \in \partial a} b_i(x_i)} \right) \prod_{i \in V} b_i(x_i)$$

  - The Gibbs free energy for b_* is given by $\mathbb{G}(b_*) = \mathbb{G}(b_a, b_i)$ and hence $\log Z = \max_{LOC(G)} \mathbb{G}(b_a, b_i)$
- G is not a Tree
  - In this case, there exists a set of marginals $\{b_i, b_a\} \in LOC(G)$ that are not marginals of any distribution b on the graph G.

The example of a case where the set $\{b_i, b_a\} \in LOC(G)$ fails to be the marginals for any distribution on G, can be obtained by considering the case of cyclic graph consisting of three variable nodes and three factor nodes as in Fig(**??**). Each variable x_i takes values in the set $\{0, 1\}$. The compatibility function for each factor node is specified in the matrix form as,

$$b_{12} = \begin{bmatrix} 0.49 & 0.01 \\ 0.01 & 0.49 \end{bmatrix} b_{23} = \begin{bmatrix} 0.49 & 0.01 \\ 0.01 & 0.49 \end{bmatrix} b_{13} = \begin{bmatrix} 0.01 & 0.49 \\ 0.49 & 0.01 \end{bmatrix}$$

From the compatibility function matrices, one observes that the factor node $(1, 2)$ prefers that variable nodes 1 and 2 be in the same state. Similarly factor node $(2, 3)$ prefers that variable nodes 2 and 3 be in the same state. On the other hand, factor node $(1, 3)$ prefers variable nodes 1 and 3 be in different states. It follows that not all of the compatibility functions can be satisfied simultaneously by a distribution. On the other hand, it can be readily checked that the marginals defined by $b_i(x_i) = (0.5, 0.5)$ and $b_{ij} = \psi_{ij}$ are locally consistent. The relation between the set of marginals of some distribution on G and the set $LOC(G)$ is summed up in Fig(**??**)

The following two results describe the relationship between the stationary points of Gibbs free energy on the set $LOC(G)$ and the fixed point of the Belief Propagation Algorithm on G and thus the relationship between the Gibbs free energy and Bethe free energy.

Proposition 7. *If the compatibility functions of a factor graph G are such that $\psi_a(\underline{x}_{\partial a}) > 0$ for all $\underline{x}_{\partial a}$ and $a \in F$, then there exists a stationary point of $\mathbb{G}(b_i, b_a)$ in the interior of $LOC(G)$. In this case, there exists a fixed point of the Belief Propagation Algorithm on G, which is given by the above stationary point.*

Claim 8. *The Bethe free energy is the Lagrangian dual of the Gibbs free energy for the locally consistent marginals .*

Proof Before writing the Lagrangian dual for the Gibbs free energy on the set $LOC(G)$, note that the constraints are specified by Eqn(8) and Eqn(9). Thus, in the dual, there will be a variable λ_i for each variable

node $i \in V$, and a variable $\lambda_{i\to a}(x_i)$ for all x_i and for each edge $(ia) \in G$. Thus the Lagrangian dual is given by

$$\mathcal{L}(\{b\},\{\lambda\}) = \mathbb{G}(b) - \sum_i \lambda_i \left(\sum_{x_i} b_i(x_i) - 1 \right) - \sum_{(ia)} \sum_{x_i} \lambda_{i\to a}(x_i) \left(\sum_{\underline{x}_{\partial a \setminus i}} b_a(\underline{x}_{\partial a}) - b_i(x_i) \right)$$

Setting $\frac{\partial \mathcal{L}}{\partial b} = 0$ leads to an expression of b as a function of λ. Substituting this into the above expression, one gets $\mathcal{L}$ as a function of just λ, which after a variable change from λ to $\{\nu_{i\to a}, \hat{\nu}_{a\to i}\}$ leads to the expression for Bethe Free energy. □

Last lecture we saw that the Bethe free energy is the Lagrange dual of the Gibbs free energy for the locally consistent marginals. Given a set of messages $\{\hat{\nu}_{a\to i}, \nu_{i\to a}\}$ we can construct the marginals by setting $\frac{\partial \mathcal{L}}{\partial b} = 0$

$$b_i(x_i) \propto \prod_{a\in\partial i} \hat{\nu}_{a\to i}(x_i)$$
$$b_a(\underline{x}_{\partial a}) \propto \psi_a(\underline{x}_{\partial a}) \prod_{i\in\partial a} \nu_{i\to a}(x_i)$$

Now by imposing the locally consistency condition we get the belief propagation fixed points. Let

$$\nu_{i\to a}(x_i) \propto \prod_{b\in\partial i\setminus a} \hat{\nu}_{a\to i}(x_i)$$

Imposing the locally consistency condition $\sum_{\underline{x}_{\partial a\setminus i}} b_a(\underline{x}_{\partial a}) = b_i(x_i)$, we get

$$\prod_{b\in\partial i} \hat{\nu}_{b\to i}(x_i) \propto \sum_{\underline{x}_{\partial a\setminus i}} \psi(\underline{x}_{\partial a}) \prod_{j\in\partial a} \nu_{j\to a}(x_j)$$
$$\hat{\nu}_{a\to i}(x_i) \prod_{b\in\partial i\setminus a} \hat{\nu}_{b\to i}(x_i) \propto \sum_{\underline{x}_{\partial a\setminus i}} \psi(\underline{x}_{\partial a}) \nu_{i\to a}(x_i) \prod_{j\in\partial a\setminus i} \nu_{j\to a}(x_j)$$
$$\hat{\nu}_{a\to i}(x_i) \propto \sum_{\underline{x}_{\partial a\setminus i}} \psi(\underline{x}_{\partial a}) \prod_{j\in\partial a\setminus i} \nu_{j\to a}(x_j)$$

6 Region-Based Approximation of the Free Energy

Region-based approximations provide a systematic scheme for constructing a hierarchy of approximate free-energy expressions. The basic idea is to decompose the system into subsystems and then approximate the free energy by combining the free energies of the subsystems. We will see that Bethe free energy is in fact a special implementation of this idea. The corresponding iterative algorithm is referred to as *generalized belief propagation*.

6.1 Regions and Region-Based Free Energy

We define a *region* R of a factor graph $G = (V, F, E)$ to be (V_R, F_R, E_R) such that

$$a \in F_R \Rightarrow \partial a \subseteq V_R$$
$$i \in V_R, a \in F_R, (i,a) \in E \Rightarrow (i,a) \in E_R$$

An example of a region is shown in Figure 3 (a). Given a region R, the Gibbs free energy of the region

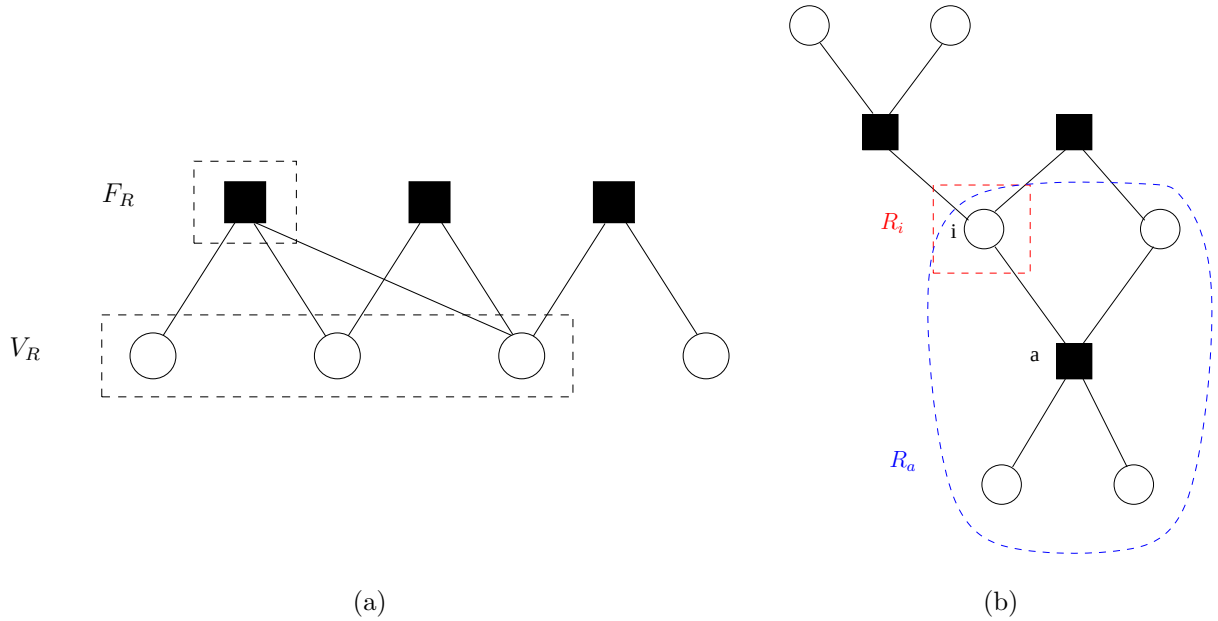


Figure 3: (a) An example of a region. (b) Regions of the Bethe approximation.

$\mathbb{G}_R : M(\mathcal{X}^{V_R}) \to \mathbb{R}$ is

$$\mathbb{G}_R[b_R] = H[b_R] + \sum_{\underline{x}_R} \sum_{a \in F_R} b_R(\underline{x}_R) \log(\psi_a(\underline{x}_{\partial a})) = H_R[b_R] - U_R[b_R]$$

where $\underline{x}_R = \{x_i : i \in V_R\}$.
The size of the domain of $\mathbb{G}_R$ is exponential in the size of the region. Therefore for small regions we can compute the associated free energy.

For a family of regions $\mathcal{R} = \{R_1, R_2, ..., R_q\}$ and the associated set of coefficients $c_\mathcal{R} = \{c_R : R \in \mathcal{R}\}$, define the *region-based free energy* $\mathbb{G}_\mathcal{R} : M(\mathcal{X}^{V_{R_1}}) \times M(\mathcal{X}^{V_{R_2}}) \times ... \times M(\mathcal{X}^{V_{R_q}}) \to R$ as

$$\mathbb{G}_\mathcal{R}\{b_\mathcal{R}\} = \sum_{R \in \mathcal{R}} c_R \mathbb{G}_R[b_R]$$

where $b_\mathcal{R} = \{b_R(\underline{x}_R) : R \in \mathcal{R}\}$.

As a check, assume that we have two disjoint factor graphs G_1 and G_2 and let $G = G_1 \bigcup G_2$. Then clearly, the free energy of G is the summation of free energies of G_1 and G_2. Now let $\mathcal{R} = \{G_1, G_2\}$, and $c_1 = c_2 = 1$. Region-based free energy, $\mathbb{G}_\mathcal{R}$, gives the exact value of the free energy. When the regions overlap then $\mathbb{G}_\mathcal{R}$ will give an approximation of the free energy.

Example 1 Bethe Free Energy
Bethe approximation is an example of region-based approximation where the regions are (Figure 3 (b))

$$\begin{aligned}
\mathcal{R} &= \{R_i, R_a : i \in V, a \in F\} \\
R_i &= (\{i\}, \emptyset, \emptyset), && \forall i \in V \\
R_a &= (\{\partial a\}, \{a\}, \{(i, a) : i \in \partial a\}), && \forall a \in F
\end{aligned}$$

and the coefficients are

$$c_i = 1 - |\partial i|, \quad \forall i \in V$$
$$c_a = 1, \qquad \forall a \in F$$

The corresponding region-based free energy is

$$\mathbb{G}_{\mathcal{R}}\{b_a, b_i\} = \sum_{i \in V}(1 - |\partial i|)H[b_i] + \sum_{a \in F}\left(H[b_a] + \sum_{\underline{x}_{\partial a}} b_a(\underline{x}_{\partial a}) \log(\psi_a(\underline{x}_{\partial a}))\right)$$

6.2 Region-Based Approximation

We approximate the free energy by

$$F = \log(Z) \approx \max_{b_{\mathcal{R}}} \mathbb{G}_{\mathcal{R}}\{b_{\mathcal{R}}\}$$

where the marginals of the regions must be consistent, i.e.

$$\forall R \supset R' : \sum_{\underline{x}_{R \setminus R'}} b_R(\underline{x}_R) = b_{R'}(\underline{x}_{R'})$$

and the coefficients must satisfy the following rules

$$\sum_{R \in \mathcal{R}} c_R \mathbb{I}(i \in V_R) = 1, \quad \forall i \in V \tag{11}$$
$$\sum_{R \in \mathcal{R}} c_R \mathbb{I}(a \in F_R) = 1, \quad \forall a \in F \tag{12}$$

Constraining the coefficients to satisfy rules (11) and (12) has the following consequences

1. If (12) holds and the marginals $\{b_R : R \in \mathcal{R}\}$ are the real marginals then $U_{\mathcal{R}}(b_{\mathcal{R}})$ is equal to the energy term of the Gibbs free energy of the real distribution, i.e, $U[\mu] = -\mathbb{E}_\mu \log(\psi(\underline{x}))$.

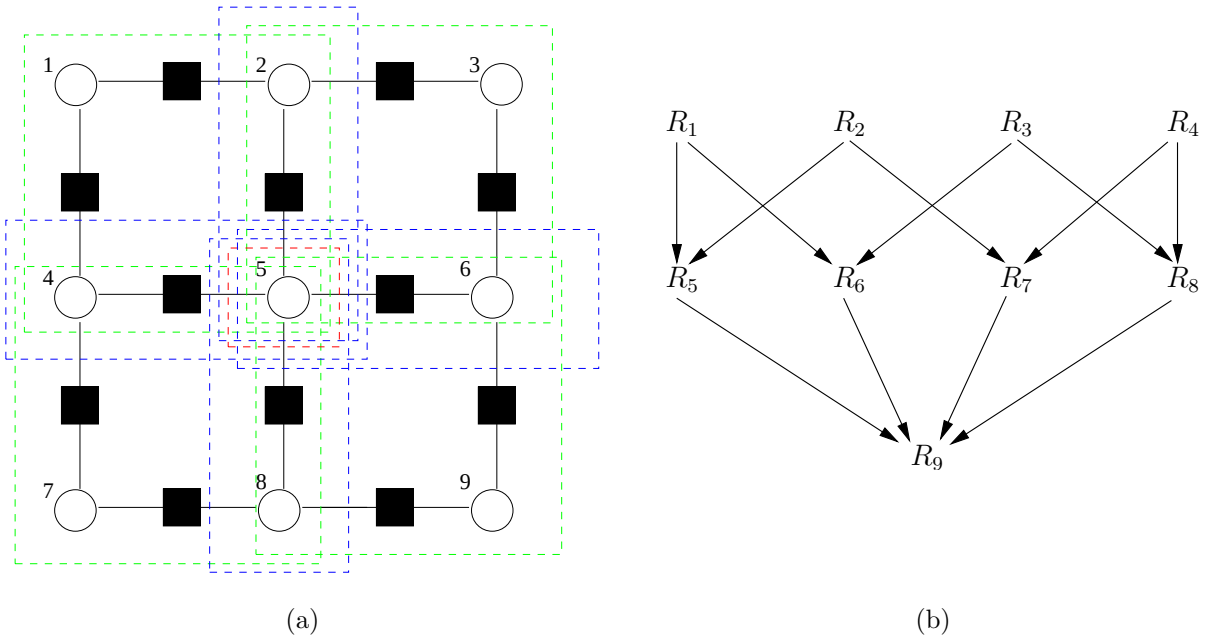
$$\begin{aligned}
U_{\mathcal{R}}(b_{\mathcal{R}}) &= -\sum_{R \in \mathcal{R}} c_R \sum_{\underline{x}_R} b_R(\underline{x}_R) \sum_{a \in F_R} \psi_a(\underline{x}_{\partial a}) \\
&= -\sum_{R \in \mathcal{R}} \sum_{a \in F} \mathbb{I}(a \in F_R) c_R \sum_{\underline{x}_{\partial a}} \mu_a(\underline{x}_{\partial a}) \psi_a(\underline{x}_{\partial a}) \\
&= -\sum_{a \in F} \left(\sum_{R \in \mathcal{R}} \mathbb{I}(a \in F_R) c_R\right) \sum_{\underline{x}_{\partial a}} \mu_a(\underline{x}_{\partial a}) \psi_a(\underline{x}_{\partial a}) \\
&= -\sum_{a \in F} \sum_{\underline{x}_{\partial a}} \mu_a(\underline{x}_{\partial a}) \psi_a(\underline{x}_{\partial a}) \\
&= U[\mu]
\end{aligned}$$

2. If (11) holds, the marginals $\{b_R : R \in \mathcal{R}\}$ are the real marginals, and the real distribution is uniform then $H_{\mathcal{R}}(b_{\mathcal{R}})$ is equal to the entropy term of the Gibbs free energy of the real distribution.

Example 2 Regions with Short Loops
Given the factor graph shown in Figure 4 (a), we define the following set of regions (and the corresponding

coefficients) where each region has at most one short loop. Regions are shown in Figure 4 (a) as dashed boxes. It is not hard to check that the coefficients satisfy rules (11) and (12).

$$
\begin{array}{llll}
V_{R_1} = \{1,2,4,5\}, & c_1 = 1 & V_{R_5} = \{2,5\}, & c_5 = -1 \\
V_{R_2} = \{2,3,5,6\}, & c_2 = 1 & V_{R_6} = \{4,5\}, & c_6 = -1 \\
V_{R_3} = \{4,5,7,8\}, & c_3 = 1 & V_{R_7} = \{5,6\}, & c_7 = -1 \\
V_{R_4} = \{5,6,8,9\}, & c_4 = 1 & V_{R_8} = \{5,8\}, & c_8 = -1 \\
V_{R_5} = \{5\}, & c_9 = 1 & &
\end{array}
$$



(a) (b)

Figure 4: (a) Factor graph and the regions. (b) Region graph corresponding to the regions.

6.3 Region Graph

Region graph $\mathcal{G}(\mathcal{R})$ of a set of regions $\mathcal{R}$ is a directed graph with vertices $R \in \mathcal{R}$ and directed edges where

$$R \to R' \Rightarrow R' \subseteq R$$

Figure 4 (b) shows a region graph corresponding to the regions of Example 2. Note that because of the condition on directed edges the graph must be a directed acyclic graph.

If there is a directed edge between R and R' then we say that R is a parent of R', $R \in P(R')$, and R' is a child of R, $R' \in C(R)$. If there is a directed path between R and R' then we say that R is an ancestor of R', $R \in A(R')$, and R' is a descendent of R, $R' \in D(R)$.

If a set of regions $\mathcal{R}$ is represented as a region graph (such representation is not necessarily unique) and the corresponding coefficients satisfy (11) and (12) and following condition

$$C_R = 1 - \sum_{R' \in A(R)} C_{R'} \quad \forall R \in \mathcal{R}$$

then the marginals of regions, $b_{\mathcal{R}}$, can be computed iteratively using Generalized Belief Propagation (GBP) algorithm.

7 Generalized Belief Propagation

To solve the following optimization problem

$$\begin{aligned} &\textit{maximaize} \quad \mathbb{G}_{\mathcal{R}}\{b_{\mathcal{R}}\} \\ &\textit{subject to} \quad \sum_{\underline{x}_{R\setminus R'}} b_R(\underline{x}_R) = b_{R'}(\underline{x}_{R'}), \quad \forall R \to R' \end{aligned}$$

We form the Lagrangian

$$\mathcal{L}(\{b_R\}, \{\lambda_{R\to R'}\}) = \mathbb{G}_{\mathcal{R}}\{b_{\mathcal{R}}\} - \sum_{R\to R'} \lambda_{R\to R'}(\underline{x}_{R'}) C_{R\to R'}(\underline{x}_{R'}))$$

where

$$C_{R\to R'}(\underline{x}_{R'}) = \sum_{\underline{x}_{R\setminus R'}} b_R(\underline{x}_R) - b_{R'}(\underline{x}_{R'})$$

Setting $\frac{\partial \mathcal{L}}{\partial b} = 0$ leads to an expression of marginals as functions of Lagrange multipliers. Now by imposing the consistency conditions, $C_{R\to R'}(\underline{x}_{R'}) = 0$, we get the update rules of the message passing algorithm.

After a variable change from $\lambda_{R\to R'}$ to $\nu_{R\to R'}$, the marginal of a region b_R is given by

$$b_R(\underline{x}_R) \propto \prod_{a\in F_R} \psi_a(\underline{x}_{\partial a}) \prod_{R_1\in P(R)} \nu_{R_1\to R}(\underline{x}_R) \prod_{R_2\in D(R)} \prod_{R_3\in P(R_2)\setminus R, D(R)} \nu_{R_3\to R_2}(\underline{x}_{R_2})$$

The relationships among R, R_1, R_2, and R_3 are described in Figure 5.

(a) (b)

Figure 5: Relationships among R, R_1, R_2, and R_3

8 Concavity of Bethe Free Energy on Trees

Two problems with Bethe Free Energy

- Bethe Free Energy is not concave

- not a bound

unless G is a tree. How do we show concavity on trees?

Helmotz Free Energy $F = \log Z = \log \sum_{\underline{x}} \prod_{a \in \mathcal{F}} \psi_a(\underline{x}_{\partial a})$. Think of this as a function of $\theta_a(\underline{x}_{\partial a})$.

$$F : \{\theta_a(\underline{x}_{\partial a}), \theta_i(x_i)\} \rightarrow F\{\theta_a, \theta_i\}$$

Substituting e^{θ} we get,

$$F(\theta) \equiv \log \{\sum_{\underline{x}} \prod_{a \in \mathcal{F}} e^{\theta_a(\underline{x}_{\partial a})} \prod_{i \in \mathcal{V}} e^{\theta_i(x_i)}\}$$

$F(\theta)$ is convex in θ as proved a few lectures before.

Consider the definition of the Bethe Free Energy

$$\mathbb{G}_{\mathrm{B}}\{b\} \quad = \quad H[b] + \mathbb{E}_b[\log(\psi(\underline{x}))]$$

We know the energy term is linear. Let's show the concavity of Bethe entropy.

Claim 9. *If G is a tree, then Bethe entropy $H_B : \{b_a, b_i\} \rightarrow H_{\mathrm{B}}\{b_a, b_i\}$ is concave.*

$$\begin{aligned} H_{\mathrm{B}}\{b\} &= \sum_{a \in F} H[b_a] - \sum_{i \in V} (1 - |\partial i|) H[b_i] \\ then, H_{\mathrm{B}}\{b\} &= \inf_{\theta} \{F\{\theta\} - \mathbb{E}_b\{\theta\}\} \\ \mathbb{E}_b\{\theta\} &= \sum_{a \in F} \sum_{\underline{x}_{\partial a}} b_a(\underline{x}_{\partial a}) \theta_a(\underline{x}_{\partial a}) + \sum_{i \in V} \sum_{x_i} b_i(x_i) \theta_i(x_i) \end{aligned}$$

Proof

$$H_{\mathrm{B}}\{b\} \quad = \quad \inf_{\theta} \{F\{\theta\} - \sum_{a \in F} \sum_{\underline{x}_{\partial a}} b_a(\underline{x}_{\partial a}) \theta_a(\underline{x}_{\partial a}) - \sum_{i \in V} \sum_{x_i} b_i(x_i) \theta_i(x_i)\}$$

Let's write the stationarity conditions

$$b_a(\underline{x}_{\partial a}) \quad = \quad \frac{\partial F(\theta)}{\partial \theta_a(\underline{x}_{\partial a})} = \mu_a^{\theta}(\underline{x}_{\partial a}) \tag{13}$$

$$b_i(x_i) \quad = \quad \mu_i^{\theta}(x_i) \tag{14}$$

where,

$$\mu_(\underline{x}^{\theta}) \equiv \frac{1}{Z} \sum_{\underline{x}} \prod_{a \in \mathcal{F}} e^{\theta_a(\underline{x}_{\partial a})} \prod_{i \in \mathcal{V}} e^{\theta_i(x_i)}$$

If G is a tree and $b \in$LOC(G), then there exists at least one set of θ's such that (1),(2) are satisfied. Further,

$$F(\theta) = H_B[\mu^{\theta}] + \mathbb{E}_{\mu}[\theta]$$

$$\Rightarrow \inf_{\theta} \{F\{\theta\} - \mathbb{E}_b\{\theta\}\} = H_B[\mu^{\theta}] + \mathbb{E}_{\mu}[\theta] - \mathbb{E}_b[\theta] = H_B[b]$$

□

9 Upper bound

For graphs that are not trees, we have

$$F(\overline{\theta}) \leq \sum_T \rho_T F(\theta_T)$$

for any collection of $\{\theta_T\}$ of parameters and weights $\{\rho_T\}$ such that

$$\sum_T \rho_T = 1, \qquad \rho_T \geq 0 \tag{15}$$

$$\sum_T \theta_T \rho_T = \overline{\theta} \tag{16}$$

Idea :

- choose θ_T such that $F(\theta_T)$ can be computed easily
- optimize over $\{\rho_T\}$, $\{\theta_T\}$ under the constraints (3),(4).

For example, we could have $T = (V_T, F_T, E_T)$ a spanning tree, where $V_T = V, F_T \subseteq F$. A tree is a spanning tree if $V_T = V$ is connected. Let's assume ρ_T is fixed and optimize over θ_T.

$$L(\{b\},\{\theta\}) = \sum_T \rho_T F(\theta_T) - \sum_{a,\underline{x}_{\partial a}} b_a(\underline{x}_{\partial a})[\sum_T \rho_T \theta_a^T(\underline{x}_{\partial a}) - \overline{\theta}_a(\underline{x}_{\partial a})] - \sum_{i,x_i} b_i(x_i)[\sum_T \rho_T \theta_i^T(x_i) - \overline{\theta}_i(x_i)]$$

Stationarity conditions with respect to θ^T gives

$$b_a(\underline{x}_{\partial a}) = \sum_T \rho_T \mu_a^{\theta_T}(\underline{x}_{\partial a}) \tag{17}$$

$$b_i(x_i) = \sum_T \rho_T \mu_i^{\theta_T}(x_i) \tag{18}$$

Further, since θ_T is non-vanishing on a tree, using the stationarity conditions we get,

$$\mathbb{G}_{\mathrm{T}}\{b\} = \sum_{a,\underline{x}_{\partial a}} b_a(\underline{x}_{\partial a})\overline{\theta}_a(\underline{x}_{\partial a}) + \sum_{i,x_i} b_i(x_i)\overline{\theta}_i(x_i) + \sum_a \rho(a)[H[b_a] - \sum_{i\in\partial a} H[b_i]] + \sum_i H[b_i]$$

where, $\rho(a)$ is the probability a factor node a belongs to a tree.

$$\rho(a) = \sum_a \sum_{T \ni a} \rho_T \in [0,1]$$

If $\rho(a)$ is close to 1, then the tree is well defined.

$\mathbb{G}_{\mathrm{T}}\{b\}$ is concave for locally consistent b_a, b_i. And we have an upper bound

$$F(\overline{\theta}) \leq \max_{b_a,b_i} \mathbb{G}_{\mathrm{T}}\{b\}$$

Notice that this depends on ρ_T only through the O(n) quantities $\rho(a)$.

$$\begin{aligned} \rho_T = \frac{1}{|T|} \to dn\rho(a) = \sum_a \rho(a) &= \sum_{a,T} \rho_T \mathbb{I}(a \in T) \\ &= \sum_T \rho_T (n-1) \\ &= n-1 \\ \to \rho(a) = \frac{n-1}{dn} & \end{aligned}$$

when $\rho(a) = 0$, $\mathbb{G}_{\mathrm{T}}$ looks like Naive Mean Field free energy.

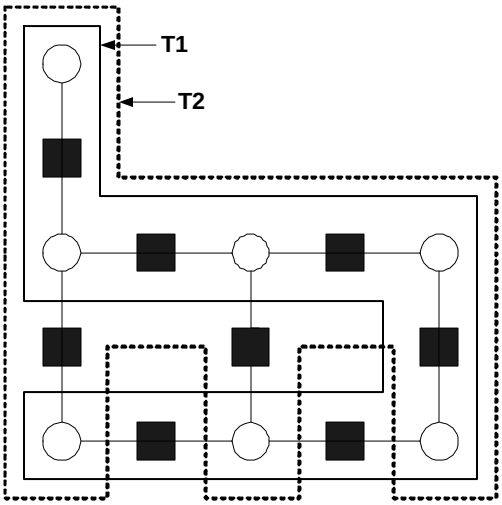


Figure 6: Example of choosing spanning trees on a graph

References

[MM09] M. Mézard and A. Montanari. *Information, Physics and Computation*. Oxford University Press, Oxford, 2009. Available online: `http://www.stanford.edu/ montanar/BOOK`.

[WJ08] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foudations and Trends in Machine Learning*, 1:1–305, 2008.

[WJW05] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky. A New Class of Upper Bounds on the Log Partition Function. *IEEE Trans. on Inform. Theory*, 51(7):2313–2335, 2005.

[YFW05] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Constructing free energy approximations and generalized belief propagation algorithms. *IEEE Trans. on Inform. Theory*, 51:2282–2313, 2005.