A non-rigorous derivation of a variational upper bound on the log-partition function in eight parts

Peter Carbonetto

Department of Computer Science
University of British Columbia
March 5, 2007

Suppose we are given some target density on the sample space $\mathcal{X}$, and it can be written in the form

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \exp\left\{\boldsymbol{\theta}^T\boldsymbol{\phi}(\boldsymbol{x}) - \Psi(\boldsymbol{\theta})\right\}, \tag{1}$$

where $\boldsymbol{\phi}(\boldsymbol{x})$ is a vector-valued function that defines the *sufficient statistics* on the sample space $\mathcal{X}$, $\boldsymbol{\theta}$ is a vector of parameters, $\boldsymbol{u}^T\boldsymbol{v}$ is the inner product of vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, and $\Psi(\boldsymbol{\theta})$ is the *log-partition function*. Its negative is the free energy commonly encounered in statistical physics [1]. The log-partition function ensures that the density sums to unity;

$$\Psi(\boldsymbol{\theta}) = \log \sum_{\boldsymbol{x}} \exp\left\{\boldsymbol{\theta}^T\boldsymbol{\phi}(\boldsymbol{x})\right\}. \tag{2}$$

In statistical physics, the density (1) is known as the Boltzmann distribution. Usually, the realizations $\boldsymbol{x}$ are vectors of random variables x_i defined at sites i, and the entries of the sufficient statistics vector factor in some fashion according to subsets of the random variables.

The first result

It can be easily shown that the first-order and second-order partial derivatives of the log-partition function with respect to individual elements of the parameter vector $\boldsymbol{\theta}$ are equal to the moments of the distribution. The partial derivatives are given by

$$\begin{aligned}
\frac{\partial\Psi}{\partial\theta_k} &= \frac{\sum_{\boldsymbol{x}} \exp\{\boldsymbol{\theta}^T\boldsymbol{\phi}(\boldsymbol{x})\}\,\phi_k(\boldsymbol{x})}{\sum_{\boldsymbol{x}} \exp\{\boldsymbol{\theta}^T\boldsymbol{\phi}(\boldsymbol{x})\}} \\
&= \frac{\sum_{\boldsymbol{x}} \exp\{\boldsymbol{\theta}^T\boldsymbol{\phi}(\boldsymbol{x})\}\,\phi_k(\boldsymbol{x})}{\exp\Psi(\boldsymbol{\theta})} \\
&= \sum_{\boldsymbol{x}} \exp\{\boldsymbol{\theta}^T\boldsymbol{\phi}(\boldsymbol{x}) - \Psi(\boldsymbol{\theta})\}\,\phi_k(x) \\
&= \sum_{\boldsymbol{x}} p(\boldsymbol{x};\boldsymbol{\theta})\,\phi_k(\boldsymbol{x}) \\
&= \mathbb{E}\{\phi_k(\boldsymbol{X})\}.
\end{aligned} \tag{3}$$

and

$$\begin{aligned}
\frac{\partial\Psi^2}{\partial\theta_k\theta_l} &= \frac{\partial}{\partial\theta_l}\sum_{\boldsymbol{x}} \exp\{\boldsymbol{\theta}^T\boldsymbol{\phi}(\boldsymbol{x}) - \Psi(\boldsymbol{\theta})\}\phi_k(\boldsymbol{x}) \\
&= \frac{\partial}{\partial\theta_l}\sum_{\boldsymbol{x}} \exp\{\boldsymbol{\theta}^T\boldsymbol{\phi}(\boldsymbol{x}) - \Psi(\boldsymbol{\theta})\}\left[\phi_l(\boldsymbol{x}) - \tfrac{\partial\Psi(\boldsymbol{\theta})}{\partial\theta_l}\right]\phi_k(\boldsymbol{x}) \\
&= \mathbb{E}\{\boldsymbol{\phi}_k(\boldsymbol{X})\boldsymbol{\phi}_l(\boldsymbol{X})\} - \mathbb{E}\{\boldsymbol{\phi}_k(\boldsymbol{X})\} \times \mathbb{E}\{\boldsymbol{\phi}_l(\boldsymbol{X})\} \\
&= \mathrm{Cov}_{p(\,\cdot\,;\boldsymbol{\theta})}\{\boldsymbol{\phi}_k(\boldsymbol{X}), \boldsymbol{\phi}_l(\boldsymbol{X})\},
\end{aligned} \tag{4}$$

where the expectations are with respect to the target $p(\boldsymbol{x};\boldsymbol{\theta})$. Since the partial derivatives (4) are covariances, the Hessian $\nabla^2\Psi$ is positive definite (the determinant of the covariance matrix must be greater than 0) and by standard results in convex analysis the log-partition function $\Psi(\boldsymbol{\theta})$ must be convex [2, Prop. B.4].

The second result

Assuming a finite sample space, the Boltzmann-Shannon entropy of the distribution (1) is defined to be

$$H(\boldsymbol{\theta}) \equiv -\sum_{\boldsymbol{x}} p(\boldsymbol{x};\boldsymbol{\theta}) \log p(\boldsymbol{x};\boldsymbol{\theta}). \tag{5}$$

Notice that the entropy is the expectation of $f(\boldsymbol{x})$ with respect to the distribution $p(\boldsymbol{x};\boldsymbol{\theta})$, where $f(\boldsymbol{x}) = \log p(\boldsymbol{x};\boldsymbol{\theta})$. Since we are always dealing with the distribution p, we assume that the entropy $H(\theta)$ is always associated with the distribution in question (1).

Now, it turns out that we can write the entropy (5) in terms of its *average energy* [9] and the log-partition function. Taking the logarithm of (1) and then operating over its expectation, we have

$$\begin{aligned}
-H(\theta) &= \mathbb{E}_{p(\,\cdot\,;\boldsymbol{\theta})}\{\log\pi(X \mid \boldsymbol{\theta})\} \\
&= \mathbb{E}_{p(\,\cdot\,;\boldsymbol{\theta})}\{\boldsymbol{\theta}^T\boldsymbol{\phi}(\boldsymbol{X}) - \Psi(\boldsymbol{\theta})\} \\
&= \mathbb{E}_{p(\,\cdot\,;\boldsymbol{\theta})}\{\boldsymbol{\theta}^T\boldsymbol{\phi}(\boldsymbol{X})\} - \Psi(\boldsymbol{\theta}),
\end{aligned} \tag{6}$$

since the log-partition function is independent of assignments $\boldsymbol{x}$. What we have in (6) is the average energy minus the log-partition function.

The third result

Following the notation of [8], the Fenchel-Legendre conjugate is a function $\Psi^\star(\boldsymbol{\mu})$ that takes as input a collection of parameters $\boldsymbol{\mu}$ and returns a number on the real line. It is defined by

$$A^\star(\boldsymbol{\mu}) \equiv \max_{\boldsymbol{\eta}} \left\{\boldsymbol{\eta}^T\boldsymbol{\mu} - \Psi(\boldsymbol{\eta})\right\}. \tag{7}$$

Also, given a set of parameters $\boldsymbol{\theta}$, we define its mean dual $\boldsymbol{\mu}$ to be

$$\boldsymbol{\mu} \equiv \mathbb{E}_{p(\,\cdot\,;\boldsymbol{\theta})}\{\boldsymbol{\phi}(\boldsymbol{X})\}. \tag{8}$$

The key result here is that if input vector $\boldsymbol{\mu}$ happens to be the dual mean of the set of parameters $\boldsymbol{\theta}$ (*i.e.* (8) is satisfied), then we have a closed-form expression for the Fenchel-Legendre conjugate function, and it is given by the Boltzmann-Shannon entropy (5). Why? The first thing to establish that the maximum of (7) is attained at point $\boldsymbol{\eta} = \boldsymbol{\theta}$ when we set $\boldsymbol{\mu}$ according to (8): denoting the function to be optimized in (7) by $Q(\boldsymbol{\eta}) \equiv \boldsymbol{\eta}^T\boldsymbol{\mu} - \Psi(\boldsymbol{\eta})$ so that $\Psi^\star(\boldsymbol{\mu}) = \max_{\boldsymbol{\eta}} Q(\boldsymbol{\eta})$, the derivative of the objective function is

$$\boldsymbol{\mu} - \mathbb{E}_{p(\,\cdot\,;\boldsymbol{\eta})}\{\boldsymbol{\phi}(\boldsymbol{X})\}.$$

from the previous result (4). Since the log-partition function is convex, it must have a unique optimum and this optimum is attained when the gradient vanishes. As a result, the optimum of (7) is attained when $\boldsymbol{\mu} = \mathbb{E}_{p(\,\cdot\,;\boldsymbol{\eta})}\{\boldsymbol{\phi}(\boldsymbol{X})\}$, which is precisely our defi-

nition (8) of the mean parameters with a new placeholder $\boldsymbol{\eta}$. So this substantiates the use of $\boldsymbol{\mu}$ in the both equation (7)—since the optimum is attained at $\boldsymbol{\eta} = \boldsymbol{\theta}$—and in (8). Putting these results together, we have

$$\begin{aligned}\Psi^\star(\boldsymbol{\mu}) &= \boldsymbol{\theta}^T\boldsymbol{\mu} - \Psi(\boldsymbol{\theta}) \qquad (9)\\ &= \boldsymbol{\theta}^T\mathbb{E}_{p(\,\cdot\,;\boldsymbol{\theta})}\{\boldsymbol{\phi}(\boldsymbol{X})\} - \Psi(\boldsymbol{\theta})\\ &= \mathbb{E}_{p(\,\cdot\,;\boldsymbol{\theta})}\boldsymbol{\theta}^T\boldsymbol{\phi}(\boldsymbol{X}) - \Psi(\boldsymbol{\theta})\\ &= -H(\boldsymbol{\theta}). \qquad (10)\end{aligned}$$

The last line follows immediately from (6).

The fourth result

The variational principle now follows almost immediately from the previous result. It should also be fairly obvious to see that we can take (9) and switch around the log partition functions[1] in order to obtain the variational representation

$$\Psi(\boldsymbol{\theta}) = \max_{\boldsymbol{\gamma}} \left\{\boldsymbol{\theta}^T\boldsymbol{\gamma} - \Psi^\star(\boldsymbol{\gamma})\right\}, \qquad (11)$$

where the domain of $\boldsymbol{\gamma}$ is the entire collection of mean parameters that correspond to legal parameterizations $\boldsymbol{\eta}$; *i.e.* the domain of interest is

$$\left\{\boldsymbol{\gamma} \,\middle|\, \exists\ \boldsymbol{\eta} \text{ such that } \boldsymbol{\gamma} = \mathbb{E}_{p(\,\cdot\,;\boldsymbol{\eta})}[\boldsymbol{\phi}(\boldsymbol{X})]\right\} \qquad (12)$$

When $\mathcal{X}$ is finite, this set can be represented by a finite collection of half-spaces [3] and is called the *marginal polytope* in [8]. In another manner of speaking: if we can compute $\Psi^\star(\boldsymbol{\gamma})$ and $\boldsymbol{\theta}^T\boldsymbol{\gamma}$ for some legal $\boldsymbol{\gamma}$, then the variational principle (11) guarantees that we have in our possession a lower bound on the log-partition function.

It is instructive to rederive this lower bound using Jensen's inequality. Jensen's inequality states that

$$f(\mathbb{E}\{X\}) \leq \mathbb{E}\{f(X)\}$$

for any convex function $f(x)$. In our case, we use the fact that $f(x) = -\log(x)$ is convex. From (10) we have

$$\begin{aligned}\Psi(\boldsymbol{\theta}) &= \log \sum_{\boldsymbol{x}} \exp\{\boldsymbol{\theta}^T\boldsymbol{\phi}(\boldsymbol{x})\}\\ &= \log \sum_{\boldsymbol{x}} \frac{p(\boldsymbol{x};\boldsymbol{\eta})}{p(\boldsymbol{x};\boldsymbol{\eta})} \exp\{\boldsymbol{\theta}^T\boldsymbol{\phi}(\boldsymbol{x})\}\\ &\geq \sum_{\boldsymbol{x}} p(\boldsymbol{x};\boldsymbol{\eta}) \log \frac{\exp\{\boldsymbol{\theta}^T\boldsymbol{\phi}(\boldsymbol{x})\}}{p(\boldsymbol{x};\boldsymbol{\eta})}\\ &= \mathbb{E}_{p(\,\cdot\,;\boldsymbol{\eta})}\boldsymbol{\theta}^T\boldsymbol{\phi}(\boldsymbol{X}) - \mathbb{E}_{p(\,\cdot\,;\boldsymbol{\eta})} \log p(\boldsymbol{X};\boldsymbol{\eta})\\ &= \boldsymbol{\theta}^T\boldsymbol{\gamma} + H(\boldsymbol{\eta})\\ &= \boldsymbol{\theta}^T\boldsymbol{\gamma} - \Psi^\star(\boldsymbol{\gamma}).\end{aligned}$$

Notice that (11) is equivalent to the classical variational principle of statistical physics [1, eq. 4.4] where both sides of (11) are negated and the maximum is replaced with a minimum.

Simply put, the variational bound (11) is not useful because we don't have an explicit representation of the Fenchel-Legendre conjugate function $A^\star$ (there are a few exceptions where we can obtain an expression for $A^\star$, but naturally these aren't very interesting problems). So [6] introduces a tractable family of approximations to the entropy, and this is what we will discuss next.

[1] This is legal because the mapping $\Psi(\boldsymbol{\theta})$ is convex, and hence invertible on its image [5, p. 56].

The fifth result

In many problems we encounter distributions of the form

$$p(x) = \frac{1}{Z}\prod_{i=1}^{n}\psi_i(x_i) \prod_{(i,j)\in E}\psi_{ij}(x_i,x_j), \qquad (13)$$

defined on an undirected graph $G = (V, E)$ with vertices $V = \{1, 2, \ldots, n\}$ and edges E. These distributions are often referred to as Markov random fields. We can rewrite this distribution in the same form as (1):

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \exp\left\{\sum_i \log\psi_i(x_i) + \sum_{(i,j)} \log\psi_{ij}(x_i,x_j) - \log Z\right\}. \qquad (14)$$

By setting $\theta = \{\theta_i \,|\, i \in V\} \cup \{\theta_{ij} \,|\, (i,j) \in E\}$, $\theta_i = \log\psi_i(x_i)$, $\theta_{ij} = \log\psi_{ij}(x_i,x_j)$ and $\Psi(\boldsymbol{\theta}) = \log Z$, we recover the original Boltzmann representation (1).

Now let's suppose that the graph G is a tree. Forgetting about the exponential family representation (14) for a brief moment, the junction tree theorem [4, Corollary 2.2] tells us that we can always write the distribution (13) in terms of its node and pairwise marginals,

$$p(\boldsymbol{x}) = \prod_i \mu_i(x_i) \prod_{(i,j)} \frac{\mu_{ij}(x_i,x_j)}{\mu_i(x_i)\mu_j(x_j)}, \qquad (15)$$

where $\mu_i(x_i)$ for all $i \in V$ and $\mu_{ij}(x_i,x_j)$ for all $(i,j) \in E$ are the local marginal distributions. (Note that the junction tree theorem applies more generally to factor graphs, which we don't discuss here.) The notation for the marginals here is not accidental—if the sufficient statistics vector $\boldsymbol{\phi}(x)$ is constructed from the site statistics x_i for all $i \in V$ and (x_i,x_j) for all $(i,j) \in E$, then the marginals are indeed the mean parameters. From now on, let's assume that this is the case: the mean parameters correspond to marginals. This is the case in the *canonical overcomplete representation*, which is just a fancy way of saying that the sufficient statistics are delta-Dirac functions $\delta_k(x_i)$, for $x_i \in \{1,\ldots,K\}$ [7].

Assuming the overcomplete representation on discrete variables and following the same derivation procedure as in (10)—that is, taking the logarithm of both sides of (15), then taking the expectation with respect to $p(\boldsymbol{x};\boldsymbol{\theta})$—we obtain

$$\begin{aligned}\mathbb{E}\left\{\log p(\boldsymbol{X};\boldsymbol{\theta})\right\} &= \textstyle\sum_i \mathbb{E}_{p(\,\cdot\,;\boldsymbol{\theta})} \log\mu_i(x_i)\\ &\quad + \textstyle\sum_{(i,j)} \mathbb{E}_{p(\,\cdot\,;\boldsymbol{\theta})} \log \frac{\mu_{ij}(x_i,x_j)}{\mu_i(x_i)\mu_j(x_j)}\\ &= \textstyle\sum_i \sum_{x_i} \mu_i(x_i) \log\mu_i(x_i)\\ &\quad + \textstyle\sum_{(i,j)} \sum_{\{x_i,x_j\}} \mu_{ij}(x_i,x_j) \log \frac{\mu_{ij}(x_i,x_j)}{\mu_i(x_i)\mu_j(x_j)},\end{aligned} \qquad (16)$$

where the expectations are all with respect to the target density. The left-hand side of (16) is $-H(\theta)$, so it is also equal to $\Psi^\star(\boldsymbol{\mu})$ when $\boldsymbol{\mu}$ is the set of marginal probabilities. Key to this derivation is the requirement that the mean parameters $\mu_i(x_i)$ be equal to the marginal probabilities (and likewise for the pairwise marginals $\mu_{ij}(x_i,x_j)$). This is not the case for other members of the exponential family (such as the Normal).[2] We will ignore this point for

[2] It is not immediately obvious how to extend the methodology here to the entire family of exponential distributions.

now, and assume an overcomplete representation on discrete random variables.

The sixth result

Suppose someone gives you a set of mean parameters $\boldsymbol{\gamma}$. That is, $\boldsymbol{\gamma}$ belongs to the marginal polytope. You then bring out another set of mean parameters $\boldsymbol{\gamma}^{\mathsf{tree}}$ that is constrained to have the same marginals, $\gamma_{ij} = \gamma_{ij}^{\mathsf{tree}}$, for all edges (i,j) belonging to the predefined spanning tree. We don't care about the marginals on the other edges not in the spanning tree. Taking the first-order Taylor expansion of the dual log-partition function $\Psi^\star$ about point $\boldsymbol{\gamma}^{\mathsf{tree}}$, we get the following lower bound on point $\boldsymbol{\gamma}$:

$$\Psi^\star(\boldsymbol{\gamma}) \geq \Psi^\star(\boldsymbol{\gamma}^{\mathsf{tree}}) + \nabla\Psi^\star(\boldsymbol{\gamma}^{\mathsf{tree}})^T(\boldsymbol{\gamma} - \boldsymbol{\gamma}^{\mathsf{tree}}), \tag{17}$$

where the gradient vector contains the partial derivatives of $\Psi^\star$ with respect to the vector $\boldsymbol{\gamma}$. We get the inequality (17) because $\Psi^\star$ is convex, and the gradient always undershoots the function surface.

Let's now suppose that $\boldsymbol{\mu}^{\mathsf{tree}}$ is the mean dual of some set of parameters $\boldsymbol{\theta}^{\mathsf{tree}}$ that projects the original parameterization $\boldsymbol{\theta}$ onto the given spanning tree; that is, $\theta_{ij}^{\mathsf{tree}} = 0$ for all edges (i,j) that do not belong to the spanning tree. Recall from (3) that the partial derivatives of the log-partition function recover the mapping between the original parameterization $\boldsymbol{\theta}$ and its mean dual $\boldsymbol{\mu}$. Due to the convexity of the log-partition function, the reverse mapping is then given by the partial derivatives

$$\frac{\partial A^\star(\boldsymbol{\gamma})}{\partial \gamma_k} = \eta_k.$$

where $\boldsymbol{\eta}$ is the mean dual of some arbitrary parameterization $\boldsymbol{\gamma}$. That means that we can rewrite the inequality (17) as

$$\begin{aligned}\Psi^\star(\mu) &\geq \Psi^\star(\boldsymbol{\mu}^{\mathsf{tree}}) + (\boldsymbol{\theta}^{\mathsf{tree}})^T(\boldsymbol{\mu} - \boldsymbol{\mu}^{\mathsf{tree}}) \\ &= \Psi^\star(\boldsymbol{\mu}^{\mathsf{tree}}) + \textstyle\sum_{(i,j)}\theta_{ij}^{\mathsf{tree}}(\mu_{ij} - \mu_{ij}^{\mathsf{tree}}),\end{aligned} \tag{18}$$

where the summation above is over all edges (i,j) in the graph. Wait…we're not quite done! Notice that the summation in (18) resolves to 0 because: 1) we've set $\theta_{ij}^{\mathsf{tree}} = 0$ for all edges (i,j) not in the spanning tree, and 2) we've already said that μ_{ij}^{tree} is equal to μ_{ij} for all edges belonging to the spanning tree. So we have the simple inequality

$$\Psi^\star(\boldsymbol{\mu}) \geq \Psi^\star(\boldsymbol{\mu}_{\mathsf{tree}}). \tag{19}$$

This inequality is true for any $\boldsymbol{\mu}$ and $\boldsymbol{\mu}_{\mathsf{tree}}$ that satisfy the conditions discussed above. This result has an intuitive interpretation. It states that entropy of the target distribution is always less than the entropy of any "moment-matched" tree-structured distribution [5, p. 214]. We caution the reader that this is not the same as matching the parameters.

The seventh result

Notice that (19) also applies for any convex combination of spanning trees. Supposing that we have a collection of trees such that a tree t is chosen with probability ρ_t, then we have

$$\Psi^\star(\gamma) \geq \sum_t \rho_t \Psi^\star(\boldsymbol{\gamma}^{(t)}). \tag{20}$$

The eighth result

We are finally ready to derive a variational upper bound on the log-partition function $\Psi(\boldsymbol{\theta})$. As we've seen, we have upper bounds on the *dual* log-partition function $\boldsymbol{A}^\star$ via a convex combination of spanning trees (20). How can we use this to achieve an upper found on the log-partition function of interest, $\Psi(\boldsymbol{\theta})$? Recall that the variational principle tells us that

$$\Psi(\boldsymbol{\theta}) = \max_{\boldsymbol{\gamma}} \left\{\boldsymbol{\theta}^T\boldsymbol{\gamma} - \Psi^\star(\boldsymbol{\gamma})\right\},$$

which is just reiterating (11). Then we plug in the bound using the convex combination (20) to obtain

$$\Psi(\boldsymbol{\theta}) \leq \max_{\boldsymbol{\gamma}} \left\{\boldsymbol{\theta}^T\boldsymbol{\gamma} - \sum_t \rho_t \Psi^\star(\boldsymbol{\gamma}^{(t)})\right\}. \tag{21}$$

We've done a bad thing here because we've ignored the role of constraints on the mean parameters $\boldsymbol{\gamma}$. Recall, they should belong to the set (12) known as the marginal polytope. Indeed, the variational bound (21) makes an implicit approximation to this set of constraints, since it only considers *local* constraints. A succinct (but lucid!) discussion can be found in Sec. 8.2 of [7].

From (10), the log-partition function $\Psi^\star(\boldsymbol{\gamma}^{(t)})$ is equal to the negative entropy $-H(\boldsymbol{\eta}^{(t)})$, where $\boldsymbol{\gamma}^{(t)}$ is the mean conjugate dual of $\boldsymbol{\eta}^{(t)}$. Also, recall that the negative entropy for a tree is special because it can be decomposed as (16). This means that we can make the following substitution in the variational bound (21):

$$\begin{aligned}\Psi^\star(\boldsymbol{\gamma}^{(t)}) = &\sum_i \sum_{x_i} \gamma_i(x_i)\log\gamma_i(x_i) \\ &- \sum_{(i,j)\in E_t}\sum_{\{x_i,x_j\}} \gamma_{ij}(x_i,x_j)\log\tfrac{\gamma_{ij}(x_i,x_j)}{\gamma_i(x_i)\gamma_j(x_j)},\end{aligned}$$

where $E_t \subseteq E$ is the set of edges present in tree t.

Alas, we are not quite done! The variational bound (21) is still not practical on its own because it involves a convex combination over the set of spanning trees, and a graph can have a *large* number of spanning trees. The final piece of the puzzle, the last ingredient to the pie, is discussed in detail in Sec. 7.2.4 of Wainwright's thesis [5] and more succinctly in Sec. III of [6]. In essence, the *spanning tree polytope* tells us that a convex combination over all mean parameters on spanning trees can be reduced to an equivalent compact representation that involves just the edge probabilities. Denoting $\tilde{\rho}_{ij}$ to be the probability that edge (i,j) appears in a spanning tree (according to the distribution ρ_t), (21) becomes

$$\begin{aligned}\Psi(\boldsymbol{\theta}) \leq \max_{\boldsymbol{\gamma}} \Bigg\{&\boldsymbol{\theta}^T\boldsymbol{\gamma} - \sum_i\sum_{x_i}\gamma_i(x_i)\log\gamma_i(x_i) \\ &- \sum_{(i,j)\in E}\tilde{\rho}_{ij}\sum_{\{x_i,x_j\}}\gamma_{ij}(x_i,x_j)\log\tfrac{\gamma_{ij}(x_i,x_j)}{\gamma_i(x_i)\gamma_j(x_j)}\Bigg\}.\end{aligned} \tag{22}$$

This is the same as Proposition 14 of [7]. Importantly, if we express the bound as $\Psi(\boldsymbol{\theta}) \leq \max_{\boldsymbol{\gamma}} Q_{\tilde{\boldsymbol{\rho}}}(\gamma)$ where the distribution of edge probabilities $\tilde{\boldsymbol{\rho}}$ is fixed, $Q_{\tilde{\boldsymbol{\rho}}}(\boldsymbol{\gamma})$ is convex and hence possesses a single maximum.

The algorithm

We now derive the tree-reweighted sum-product updates which comprise an algorithm for coordinate ascent for computing the optimum of to (22). We have a constrained optization problem, since our approximation to the marginal polytope (which we haven't dis-

cussed) requires that

$$\sum_{x_i} \gamma_{ij}(x_i, x_j) = \gamma_j(x_j) \qquad \text{and} \qquad \sum_{x_j} \gamma_{ij}(x_i, x_j) = \gamma_i(x_i)$$

for all edges $(i, j) \in E$. For the functions $\gamma_i(x_i)$ and $\gamma_{ij}(x_i, x_j)$ and to be valid mean parameters, we also require that they sum to one over their arguments. To fully specify the simplex, one generally includes inequality constraints to ensure that the beliefs are strictly positive [9]. We will assume that all the mean parameters are strictly positive, in which case the corresponding Lagrange multipliers will vanish from the Lagrangian function. Incorporating the equality constraints discussed above, the Langrangian function [2, Sec. 3.1.3] for our optimization problem is

$$\begin{aligned} L(\boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{\alpha}) = & \sum_i \sum_{x_i} \gamma_i(x_i)\theta_i(x_i) \\ & + \sum_{(i,j)} \sum_{x_i} \sum_{x_j} \gamma_{ij}(x_i, x_j)\theta_{ij}(x_i, x_j) \\ & + \sum_i (w_i - 1) \sum_{x_i} \gamma_i(x_i) \log \gamma_i(x_i) \\ & - \sum_{(i,j)} \tilde{\rho}_{ij} \sum_{x_i} \sum_{x_j} \gamma_{ij}(x_i, x_j) \log \gamma_{ij}(x_i, x_j) \\ & + \sum_i \sum_{j \in N(i)} \sum_{x_i} \lambda_{ji}(x_i) \left\{ \sum_{x_j} \mu_{ij}(x_i, x_j) - \mu_i(x_i) \right\} \\ & + \sum_i \alpha_i \left\{ 1 - \sum_{x_i} \gamma_i(x_i) \right\} \\ & + \sum_{(i,j)} \alpha_{ij} \left\{ 1 - \sum_{x_i} \sum_{x_j} \gamma_{ij}(x_i, x_j) \right\}, \end{aligned} \tag{23}$$

where we define $w_i = \sum_{j \in N(i)} \tilde{\rho}_{ij}$ and denote $N(i) = \{j \mid (i, j) \in E\}$ to be the set of neighbours of vertex i. The scalars $\lambda_{ij}(x_j)$ and $\lambda_{ji}(x_i)$ are the Lagrange multipliers associated with the marginalization constraints and α_i and α_{ij} are multipliers that enforce the normalization constraints. Since all the constraints are linear, every optimum $\boldsymbol{\gamma}$ possesses Lagrange multipliers, even when $\boldsymbol{\gamma}$ is not regular [2]. Note the similarity to the Bethe free energy [9], the main difference being that entropy terms are weighted by the edge probabilities.

We start by taking partial derivatives of (23) with respect to the singleton and pairwise mean parameters. We have

$$\begin{aligned} \frac{\partial L}{\partial \gamma_i(x_i)} = & \theta_i(x_i) + w_i \log \gamma_i(x_i) + w_i \\ & - \textstyle\sum_{j \in N(i)} \lambda_{ji}(x_i) + \alpha_i \end{aligned} \tag{24}$$

$$\begin{aligned} \frac{\partial L}{\partial \gamma_{ij}(x_i, x_j)} = & \theta_{ij}(x_i, x_j) - \tilde{p}_{ij} \log \gamma_{ij}(x_i, x_j) - \tilde{p}_{ij} \\ & + \lambda_{ij}(x_j) + \lambda_{ji}(x_i) + \alpha_{ij}. \end{aligned} \tag{25}$$

Setting the partial derivatives to zero, we obtain the correct sum-product updates.

References

[1] S. M. Aji and R. J. McEliece. The Generalized distributive law and free energy minimization. In *Proceedings of the 39th Allerton Conference*, pages 672–681, 2001.

[2] D. P. Bertsekas. *Nonlinear programming*. Athena Scientific, second edition, 1999.

[3] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[4] M. A. Paskin. *Exploiting locality in probabilistic inference*. PhD thesis, 2004.

[5] M. J. Wainwright. *Stochastic processes on graphs with cycles: geometric and variational approaches*. PhD thesis, 2002.

[6] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky. A New class of upper bounds on the log partition function. *IEEE Trans. Inform. Theory*, 51:2313–2335, 2005.

[7] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. Technical report, EECS Dept., University of California, Berkeley, 2003.

[8] M. J. Wainwright and M. I. Jordan. Variational inference in graphical models: the view from the marginal polytope. In *Proceedings of the 41st Allerton Conference*, 2003.

[9] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, 51:2282–2312, July 2005.