Approximate Inference in Boltzmann Machines

Max Welling *Yee Whye Teh*

Department of Computer Science
University of Toronto
6 King's College Road
Toronto, M5S 3G5 Canada

Keywords: Advanced Mean Field Methods, Loopy Belief Propagation, Inference, Boltzmann Machines

Abstract

Inference in Boltzmann machines is NP-hard in general. As a result approximations are often necessary. We discuss first order mean field and second order Onsager truncations of the Plefka expansion of the Gibbs free energy. The Bethe free energy is introduced and rewritten as a Gibbs free energy. From there a convergent *belief optimization* algorithm is derived to minimize the Bethe free energy. An analytic expression for the linear response estimate of the covariances is found which is exact on Boltzmann trees. Finally, a number of theorems is proven concerning the Plefka expansion, relating the first order mean field and the second order Onsager approximation to the Bethe approximation. Experiments compare mean field approximation, Onsager approximation, belief propagation and belief optimization.

1 Introduction

In 1982, Hopfield showed that a network of symmetrically-coupled binary threshold units has a simple quadratic energy function that governs its dynamic behavior [6]. When the nodes are deterministically updated one at a time the network settles to an energy minimum and Hopfield suggested using these minima to store content-addressable memories. Hinton and Sejnowski realised that the energy function can be viewed as an indirect way of defining a probability distribution over all the binary configurations of the network and that if the right stochastic updating rule is used, the dynamics eventually produces samples from this Boltzmann distribution [4] [5].

If this "Boltzmann machine" is divided into a set of visible nodes whose states are clamped at the data and a disjoint set of hidden nodes, the stochastic updating produces samples from the posterior distribution over configurations of the hidden nodes given the current data. Hinton and Sejnowski suggested that the sampled hidden configurations could be viewed as perceptual interpretations of the observed data in terms of hidden features. They also showed that there is a surprisingly simple algorithm for performing maximum likelihood learning of the weights that define these hidden features. The simplicity and locality of this learning rule led to much interest, but the settling time required to get samples from the right distribution and the high noise in the estimates required for the learning rule made learning slow and unreliable.

There is however a number of approximate methods which can be employed, among which the "mean field" approximation is best known [23]. There, the best distribution is sought that assumes independence among all the nodes. Variational approximations which employ more structured but still tractable posterior distributions (e.g. chains or trees) have been proposed to improve on the simple independence assumption of mean field [7].

From a different perspective, the mean field free energy can also be viewed as the first term in a series expansion around small weights where the means are kept fixed (Plefka's expansion) [20]. Taking into account the next order in this expansion produces Onsager's reaction term. Higher orders have also been computed in the physics literature [31] (for additional information see [13],[21],[16],[18]).

A third possibility which has received a lot of attention lately in the AI community is belief propagation. It is an efficient local message passing protocol for exact inference on trees [22]. Applying the same rules to graphs with cycles (loopy belief propagation) has proven a successful strategy for approximate inference [15]. In particular, it was shown that the celebrated method of "turbo decoding" is equivalent to loopy belief propagation on an appropriate graphical model [12],[2].

At first there was not much theoretical justification for applying belief propagation to loopy graphs. Lately however, much progress has been made in understanding the convergence properties of the algorithm and the quality of the approximation [27],[26]. The most significant breakthrough came with the insight that the fixed points of belief propagation are actually the stationary points of the Bethe free energy. It not only clarified the nature of the approximation but it also opened the way to the more sophisticated Kikuchi approximations and an algorithm to solve for its stationary points [32].

In this paper we will prove a number of theorems, some of which were conjectured in the physics literature a decade ago, which clarify the relation between the Bethe approximation and the Plefka expansion of the exact Gibbs free energy. We also propose a novel algorithm, named belief optimization, to minimize the Bethe free energy directly, as an alternative to the fixed point equations of belief propagation. Unlike belief propagation, this algorithm is provably convergent. Moreover a new linear response estimate is derived to compute the covariances between all pairs of nodes and show that it is exact on Boltzmann trees. Some of the results in this paper were also reported in [29].

In experiments we confirm that belief optimization and belief propagation give identical results when the latter converges. We also show that in the cases when belief propagation can not be made to converge, the Bethe free energy is likely to be a bad approximation and belief optimization will also give inaccurate results. Finally, the experiments confirm that the linear response estimates of the covariances in the Bethe approximation are better than their counterparts in the mean field and Onsager approximations.

2 Model and Notation

The model under consideration, a Boltzmann machine, can be represented as an undirected graphical model, with binary nodes taking values either 0 or 1. Some nodes may be directly observed and are denoted by v_k, while others remain unobserved (hidden) and will be denoted by h_i. The probability function is defined through its energy as

$$P(\{v_k\},\{h_i\}) = \frac{1}{\mathcal{Z}} \exp\left(-E(\{v_k\},\{h_i\})\right) \tag{1}$$

where $\mathcal{Z}$ is the normalization constant (the "partition function"). The energy contains bias terms, with thresholds θ_i for the hidden units and α_k for the visible units. Pairwise interaction terms are defined through symmetric weights $W_{ij} = W_{ji}$ between the hidden units, $V_{kl} = V_{lk}$ between the visible units and J_{ik} between hidden and visible units. There are no self interactions, i.e. $W_{ii} = V_{kk} = 0$. The total energy can therefore be written as,

$$E(\{v_k\},\{h_i\}) = -\sum_{(kl)} V_{kl} v_k v_l - \sum_k \alpha_k v_k - \sum_{(ij)} W_{ij} h_i h_j - \sum_i \theta_i h_i - \sum_{(ik)} J_{ik} h_i v_k \tag{2}$$

where (ij) denote all pairs of neighboring hidden nodes and similarly for (kl) and (ik) (k and l index visible nodes). In this paper we will only be concerned with *inference*, i.e. our ability to compute the posterior probability function

$$P(\{h_i\}|\{v_k\} = \{d_k\}) = \frac{1}{Z}\exp\left(-E(\{h_i\},\{d_k\})\right) \tag{3}$$

where $\{d_i\}$ is a data-vector and Z is the partition function for the posterior distribution:

$$Z = \sum_{\{h_i\}} \exp(-E(\{h_i\},\{d_k\})) \tag{4}$$

The energy for the posterior distribution is given by,

$$E(\{h_i\},\{d_k\}) = -\sum_{(ij)} W_{ij} h_i h_j - \sum_i (\theta_i + \sum_{k\in N_v(i)} J_{ik} d_k) h_i \tag{5}$$

where $N_v(i)$ denotes the set of visible neighbors of node i.

The effect of the observation is to change the thresholds by an amount $\theta_i \to \theta_i + \sum_{k\in N_v(i)} J_{ik} d_k$. In the following we will simply absorb these shifts into the definitions of the thresholds. The problem of computing the posterior is now equivalent to the computation of the "prior" $P(\{h_i\})$ with these shifted thresholds (and visible nodes removed). In subsequent sections we will concern ourselves with computing this prior $P(\{h_i\})$. Further, instead of trying to compute the entire prior probability table for all possible states $\{h_i\}$, we will restrict ourselves to approximating the marginals $p_i(h_i)$ and pairwise marginals $p_{ij}(h_i, h_j)$.

For binary variables it is convenient to reparametrize these marginals as follows,

$$p_i(h_i = 1) = \langle h_i \rangle = q_i \tag{6}$$

$$p_{ij}(h_i = 1, h_j = 1) = \langle h_i h_j \rangle = \xi_{ij} \tag{7}$$

All the other entries of the single node and pairwise probability tables can be expressed in terms of this set of independent parameters,

$$\begin{aligned} p_{ij}(h_i = 1, h_j = 0) &= q_i - \xi_{ij} && (8) \\ p_{ij}(h_i = 0, h_j = 1) &= q_j - \xi_{ij} && (9) \\ p_{ij}(h_i = 0, h_j = 0) &= \xi_{ij} + 1 - q_i - q_j && (10) \\ p_i(h_i = 0) &= 1 - q_i && (11) \end{aligned}$$

It can also be checked that all marginalization constraints are satisfied, e.g.

$$\begin{aligned} \sum_{h_i=0,1} p_{ij}(h_i, h_j = 1) &= q_j && (12) \\ \sum_{h_i=0,1} p_{ij}(h_i, h_j = 0) &= 1 - q_j && (13) \\ \sum_{h_i,h_j=0,1} p_{ij}(h_i, h_j) &= 1 && (14) \end{aligned}$$

Obviously, there are certain constraints on the values of the $\{q_i\}$ and the $\{\xi_{ij}\}$ to ensure that all probabilities are between 0 and 1, but they are easy to handle as will become evident in the rest of this paper.

3 The Mean Field Approximation

In the mean field (MF) approximation we try to find a factorized distribution that best describes the true posterior distribution. The most general factorized distribution for binary variables has the form,

$$Q^{\mathrm{MF}}(\{h_i\}) = \prod_i q_i^{h_i} \, (1 - q_i)^{1-h_i} \tag{15}$$

The variational parameters $\{q_i\}$ represent the means $q_i = \langle h_i \rangle$ and are chosen so that $Q^{\mathrm{MF}}(\{h_i\})$ is close to the true posterior by minimizing the following Kullback-Leibler (KL) divergence,

$$\mathrm{KL}\left(Q^{\mathrm{MF}}(\{h_i\}) || P(\{h_i\})\right) \tag{16}$$

Using the explicit expressions for Q^{MF} (eqn.15) and P (eqn.3) this can be written as,

$$\begin{aligned} \mathrm{KL}\left(Q^{\mathrm{MF}}||P\right) &= \langle E\rangle_{Q^{\mathrm{MF}}} - S(Q^{\mathrm{MF}}) + \log(Z) \qquad (17) \\ \langle E\rangle_{Q^{\mathrm{MF}}} &= -\sum_{(ij)} W_{ij} q_i q_j - \sum_i \theta_i q_i \\ S(Q^{\mathrm{MF}}) &= -\sum_i \left(q_i \ln(q_i) + (1-q_i)\ln(1-q_i)\right) \qquad (18) \end{aligned}$$

where $-\log(Z)$ is the exact free energy. It is now an easy matter to derive the mean field equations by taking the gradient of the above expression with respect to q_i,

$$\frac{\partial \mathrm{KL}}{\partial q_i} = -\sum_{j\in N(i)} W_{ij} q_j - \theta_i + \log\left(\frac{q_i}{1-q_i}\right) \qquad (19)$$

with $N(i)$ the neighbors of node i, and equating them to zero to get

$$q_i = \sigma\left(\sum_{j\in N(i)} W_{ij} q_j + \theta_i\right) \qquad (20)$$

where $\sigma(x) = 1/(1+\exp(-x))$ is the sigmoid function. When these mean field equations are run sequentially, i.e. we fix all means q_j except q_i over which we minimize, the KL-divergence is convex in q_i and eqn.20 finds the minimum in one step. This procedure can thus be interpreted as coordinate descent in the $\{q_i\}$ and each step is guaranteed to decrease the KL-divergence. However, this procedure could suffer from slow convergence or entrapment in local minima. Alternatively, all parameters q_i can be updated in parallel, which does not have the guarantee of decreasing the cost-function at every iteration, but may converge faster. In practice, one often observes oscillatory behavior which can be counteracted by damping the updates eqn. 20.[1]. Finally, one can use any gradient based optimization technique to minimize over all the nodes $\{q_i\}$ simultaneously, making sure all $\{q_i\}$ remain between 0 and 1. This can for instance be achieved by reparametrizing $q_i = \sigma(y_i)$ and

[1] $q_i \leftarrow \alpha q_i + (1-\alpha)\sigma(\cdot) \quad \alpha \in [0,1)$

performing gradient descent on y_i using,

$$\frac{\partial \mathrm{KL}}{\partial y_i} = \frac{\partial \mathrm{KL}}{\partial q_i}\, q_i\, (1 - q_i) \tag{21}$$

We will see in subsequent sections that these three options: coordinate descent, fixed point equations, and gradient descent have analogous counterparts in the Onsager and Bethe approximations.

It is important to understand when this mean field approximation is expected to be accurate. For large, densely connected, weakly interacting systems the cumulative effect of all nodes behaves as a "rigid" (mean) field, $\sum_{j \in N(i)} W_{ij} h_j \approx H_i$ which acts as an additional bias term, resulting in a factorized distribution. Also, the factorized MF distribution is clearly unimodal, and could therefore never represent multi-modal posterior distributions accurately. In particular, the KL-divergence eqn. 17, with Q in its first argument, penalizes states with small posterior probability but non-vanishing probability under the MF distribution much harder than the other way around. The result of this asymmetry in the KL-divergence is that the MF distribution will choose to represent only one mode, ignoring the other ones. A typical situation where we expect multiple modes in the posterior is when there is not a lot of evidence clamped on the observation nodes. Consider for instance the situation when the thresholds are given by $\theta_i = -\frac{1}{2}\sum_j W_{ij}$ in which case there is a symmetry in the system where changing all nodes by $h_i \to 1 - h_i$ leaves all probabilities invariant. This implies that there are at least two modes. In general, we expect many more modes, and the MF distribution can only capture one. Moreover, when the interactions are strong, we expect these modes to be concentrated on one state, with little fluctuation around them. The marginals predicted by MF would therefore be close to either 1 or 0 (they are polarized), while the true marginal posterior probabilities are $\frac{1}{2}$ due to the symmetry.

One way to overcome some of the difficulties mentioned above is to use more structured variational distributions Q and minimize again the KL-divergence eqn.17 [7, 30]. We will however pursue a different approach in the following, where we directly approximate the *free energy* of the system, without making reference to a variational distribution Q.

3.1 The Mean Field Approximation - Linear Response

Although the mean field approximation assumes independence between the variables $\{h_i\}$, it is still possible to obtain a nontrivial estimate of the correlations within the MF framework [9]. The idea is to exploit the fact that the negative log-partition function is the generating function of the centralized moments (or cumulants), i.e.

$$\langle h_i \rangle = \frac{\partial}{\partial \theta_i} \log(Z) \tag{22}$$

$$C_{ij} = \langle h_i h_j \rangle - \langle h_i \rangle \langle h_j \rangle = \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log(Z) = \frac{\partial \langle h_i \rangle}{\partial \theta_j} \tag{23}$$

The linear response estimate for the correlations is obtained by replacing the true $\langle h_i \rangle$ with the approximate q_i and inverting the partial derivatives matrix:

$$[C_{ij}] = \left[\frac{\partial \langle h_i \rangle}{\partial \theta_j}\right] \approx \left[\frac{\partial q_i}{\partial \theta_j}\right] = \left[\frac{\partial \theta_j}{\partial q_i}\right]^{-1} \tag{24}$$

where $[A_{ij}]$ denotes a matrix with elements A_{ij}. The mean field equations eqn.20 provide us with an expression relating $\{\theta_j\}$ to $\{q_i\}$ upon convergence. By taking derivatives of eqn.20 with respect to q_i, solving for $\partial\theta_j/\partial q_i$, and plugging into eqn.24, we get

$$[C_{ij}] \approx \left[\frac{\partial \theta_j}{\partial q_i}\right]^{-1} = \left[\frac{\delta_{ji}}{q_j(1-q_j)} - W_{ji}\right]^{-1} \tag{25}$$

In the next section we will see that the mean field approximation can also be viewed as the first order approximation in a small weight expansion of the (Gibbs) free energy. In that context and for future reference we will expand the expression for the correlations up to linear order in the weights, giving

$$C_{ij} \approx q_i(1-q_i)\delta_{ij} + W_{ij} q_i(1-q_i) q_j(1-q_j) \tag{26}$$

4 The Legendre Transform and Plefka's Expansion

In the previous section the approach to obtain approximate posterior marginals was to define a variational distribution Q whose KL-divergence with the true posterior P was minimized over a set of free parameters $\{q_i\}$. The KL-divergence was written as (eqn.17),

$$KL(Q||P) = F(Q) - F(P); \tag{27}$$

where

$$F(Q) = \langle E \rangle_Q - S(Q) \tag{28}$$

is the *variational free energy* and $F(P) = -\log(Z)$ is the true free energy. Since the variational free energy and the KL-divergence are equal up to a constant, we can also interpret the MF procedure in the previous section as trying to minimize the variational free energy over distributions Q. Equation 27 is then saying that since the KL-divergence is non-negative, the true free energy is always upper bounded by the variational free energy. If we leave Q unconstrained (except for the fact that it should sum to one), then it is easy to show that the minimizing distribution is precisely the Boltzmann (or equilibrium) distribution,

$$Q^{\mathrm{EQ}}(\{h_i\}) = \frac{1}{Z} \exp\left(-E(\{h_i\})\right) \tag{29}$$

We may also choose to perform a partial, constrained minimization over Q, where we keep the marginals fixed. The result is what is known as the Gibbs free energy in the physics literature,

$$G(\{q_i\}) = \min_Q \ \{F(Q) \ \mid \ \langle h_i \rangle_Q = q_i\} \tag{30}$$

The rationale behind this partial minimization is that we can approximate this Gibbs free energy in terms of a small weight expansion, known as the Plefka expansion. The approximate expression for the Gibbs energy should then still be minimized over the parameters $\{q_i\}$ to obtain an approximation to the true free energy which is unfortunatly no longer guaranteed to be an upper bound.

In the following we will briefly describe this Plefka expansion without going into details. For more background material see [20],[3] and [24]. The natural way to include the constraints on the marginals is by introducing Lagrange multipliers $\{\lambda_i\}$, and adding the following term to the variational free energy,

$$F(Q) \to F(Q) - \sum_i \lambda_i \ (\langle h_i \rangle_Q - q_i) \tag{31}$$

where the Lagrange multipliers $\{\lambda_i\}$ should be chosen so as to enforce the constraints. One can now minimize $F(Q)$ over Q in terms of the $\{\lambda_i\}$ and the $\{q_i\}$. The solution is again a Boltzmann distribution eqn.29, but with a modified energy which includes additional bias terms,

$$E(\{h_i\}) \to E(\{h_i\}) - \sum_i \lambda_i h_i \tag{32}$$

After inserting this expression back into the variational free energy, we can find the values of the Lagrange multipliers $\{\lambda_i\}$ as a function of the $\{q_i\}$ by maximizing over them. The final result is then,

$$G(\{q_i\}) = \max_{\{\lambda_i\}} \left\{ \sum_i \lambda_i q_i - \log\left(Z(\{\lambda_i\})\right) \right\} \tag{33}$$

where $Z(\{\lambda_i\})$ is the normalizing constant for the Gibbs distribution with energy defined in eqn.32. Equation 33 is known as the Legendre transform between $\{\lambda_i\}$ and $\{q_i\}$. By shifting the Lagrange multipliers as follows,

$$\lambda_i' = \lambda_i + \theta_i \tag{34}$$

we can pull the contribution of the thresholds to the Gibbs free energy out of the Legendre transform,

$$G(\{q_i\}) = -\sum_i \theta_i q_i + \max_{\{\lambda_i\}'} \left\{ \sum_i \lambda_i' q_i - \log\left(Z'(\{\lambda_i'\})\right) \right\} \tag{35}$$

where Z' is the partition function with all thresholds $\{\theta_i\}$ set to zero,

$$Z'(\{\lambda_i'\}) = \sum_{\{h_i\}} \exp\left(-\sum_{(ij)} W_{ij} h_i h_j - \sum_i \lambda_i' h_i \right) \tag{36}$$

Plefka's expansion can be derived by Taylor expanding the Gibbs free energy for small weights W_{ij}[2]. The approximate Gibbs free energy is then obtained by truncating this series expansion, and in lowest order this turns out to be the MF approximation.

[2]Notice that in eqn.33 both $\log(Z)$ as well as $\{\lambda_i\}$ depend on W_{ij} and should be expanded out.

5 Onsager's Second Order Reaction Term

Truncating the Plefka expansion after the second order results in what we will call the Onsager approximation,given by

$$\begin{aligned} G^{\text{ONSAGER}}(\{q_i\}) = & -\sum_{(ij)} W_{ij} q_i q_j - \sum_i \theta_i q_i + \sum_i q_i \ln(q_i) + (1-q_i)\ln(1-q_i) \\ & -\frac{1}{2}\sum_{(ij)} W_{ij}^2 \; q_i(1-q_i) q_j (1-q_j) \end{aligned} \tag{37}$$

Higher order contributions were computed in [3]. The extra term in eqn.37, takes into account dependencies between nodes which were ignored in the MF approximation. Onsager's term was first proposed in an entirely different context for a specific physical system valid only under strict conditions [17]. In this paper it will be considered as just another order in the Plefka expansion of some finite spin system (i.e. a Boltzmann machine). Hence, we cannot expect it to improve the accuracy of the approximation under general conditions. The expansion only makes sense when it actually exists (i.e. converges), which implies that the weights must be small. When not applied with care the extra term could actually deteriorate the MF approximation [10]. To compute the q_i we simply compute their derivatives again,

$$\frac{\partial G^{\text{ONSAGER}}}{\partial q_i} = -\sum_{j \in N(i)} W_{ij} q_j - \theta_i + \log\left(\frac{q_i}{1-q_i}\right) - \frac{1}{2}(1-2q_i) \sum_{j\in N(i)} W_{ij}^2 \; q_j(1-q_j) \tag{38}$$

Equating the derivatives to zero provides a set of coupled fixed point equations which are generalizations of the MF equations,

$$q_i = \sigma\left(\sum_{j\in N(i)} W_{ij} q_j + \theta_i + \frac{1}{2}(1-2q_i)\sum_{j\in N(i)} W_{ij}^2 \; q_j(1-q_j)\right) \tag{39}$$

Unfortunately, running these equations sequentially (or in parallel) does not guarantee that the Onsager-Gibbs free energy decreases, since q_i appears both on the left and the right hand side. We may again apply damping to encourage convergence. It is however not hard to see that if we fix the means q_j of all the neighbors of a node i, the Onsager-Gibbs free energy is convex in the mean q_i of this central node. We can therefore employ any standard minimizer to find this unique minimum and cycle through the nodes. This procedure can again be interpreted as a coordinate descent algorithm and every step is guaranteed to decrease the Onsager-Gibbs free energy. Alternatively, one may wish to perform gradient descent on all the means simultaneously while making sure their values stay within $[0, 1]$.

5.1 The Onsager Approximation - Linear Response

By applying the linear response estimate for the covariances to the Onsager-Gibbs free energy we can improve the MF approximation by one order in the weights W_{ij},

$$\begin{aligned}
[C_{ij}] \approx \quad & \left[\frac{\delta_{ji}}{q_j(1-q_j)} - W_{ji} + \sum_k W_{jk}^2 q_k(1-q_k)\ \delta_{ji} - \frac{1}{2}W_{ji}^2(1-2q_j)(1-2q_i)\right]^{-1} \\
\approx \quad & \Big[q_i(1-q_i)\delta_{ij} + W_{ij} q_i(1-q_i) q_j(1-q_j) + \\
& q_i(1-q_i)q_j(1-q_j)\sum_k W_{ik}W_{kj}\ q_k(1-q_k) - \\
& q_i(1-q_i)q_j(1-q_j)\sum_k W_{ik}^2\ q_k(1-q_k)\delta_{ij} + \\
& q_i(1-q_i)q_j(1-q_j)\frac{1}{2}W_{ij}^2(1-2q_i)(1-2q_j)\Big]
\end{aligned} \tag{40}$$

The first two terms clearly correspond to the MF-linear response result.

6 The Bethe Approximation

The Bethe approximation first made its appearance in the field of approximate inference and error correcting decoding in [19] and [8] under the names TAP approximation and cavity method, but has a much longer history in physics [13],[21]. The relation between belief propagation and the Bethe approximation was further clarified in [32], where it was shown that belief propagation, even when applied to loopy graphs, has fixed points at the stationary points of the Bethe free energy. In this section we will define the

Bethe free energy for Boltzmann machines, rederive belief propagation update rules and describe a novel algorithm that directly minimizes this Bethe free energy.

In the case of the Onsager and MF approximation we first defined the Gibbs free energy through a partial minimization (equivalent to a Legendre transform) after which this was approximated with an expansion (Plefka's expansion) around small weights. We will follow a similar procedure for the Bethe approximation, by formally defining a more constrained Gibbs free energy, and then proposing an approximation of this Gibbs free energy which is called the Bethe free energy (see [31]). This more constrained Gibbs free energy is defined by,[3]

$$\mathcal{G}(\{q_i, \xi_{ij}\}) = \min_{Q} \ \{F(Q) \ \mid \ \langle h_i \rangle_Q = q_i \ \ \& \ \ \langle h_i h_j \rangle_Q = \xi_{ij}\} \tag{41}$$

We have denoted this more constrained free energy with $\mathcal{G}(\{q_i, \xi_{ij}\})$, to distinguish it from the Gibbs free energy $G(\{q_i\})$ which is the result of minimizing $\mathcal{G}(\{q_i, \xi_{ij}\})$ over $\{\xi_{ij}\}$.

We will now proceed to define the Bethe free energy $\mathcal{G}^{\text{BETHE}}$ and in later subsections convert it to a Bethe-Gibbs free energy by minimizing it over the correlations $\{\xi_{ij}\}$. Since the energy of a Boltzmann machine is a quadratic function of the states its average in terms of $\{q_i\}$ and $\{\xi_{ij}\}$ is simple and exact,

$$E = -\sum_{(ij)} W_{ij}\xi_{ij} - \sum_i \theta_i q_i \tag{42}$$

The approximation is made for the entropy term of the free energy. The idea is that we want to correct the MF approximation which overestimates the entropy due to its assumption that all nodes are independent. The natural next step is to take pairwise dependencies into account. But just adding all pairwise entropy contributions to the MF approximation would clearly over-count the entropy contributions at the nodes. Correcting for this over-

[3] Note that for particular settings of $\{q_i, \xi_{ij}\}$ there is no distribution Q with $\langle h_i \rangle_Q = q_i$ *and* $\langle h_i h_j \rangle_Q = \xi_{ij}$ [11]. This is unimportant for the development of this paper and for consistency we let $\mathcal{G}(\{q_i, \xi_{ij}\}) = \infty$ in such cases.

counting then gives the following approximation to the entropy,

$$\begin{aligned} S^{\text{BETHE}} &= \sum_i S_i + \sum_{(ij)} (S_{ij} - S_i - S_j) \\ &= \sum_i (1 - z_i) S_i + \sum_{(ij)} S_{ij} \end{aligned} \tag{43}$$

where z_i is defined as the number of neighbors of node i, S_i is the mean field entropy for node i,

$$S_i = -(q_i \ln(q_i) + (1 - q_i) \ln(1 - q_i)) \tag{44}$$

and S_{ij} is the pairwise entropy which can be written as,

$$\begin{aligned} S_{ij} = -(\; & \xi_{ij} \ln(\xi_{ij}) + \; (\xi_{ij} + 1 - q_i - q_j) \ln(\xi_{ij} + 1 - q_i - q_j) \\ & + (q_i - \xi_{ij}) \ln(q_i - \xi_{ij}) + \; (q_j - \xi_{ij}) \ln(q_j - \xi_{ij})) \end{aligned} \tag{45}$$

The Bethe free energy is now defined in terms of the energy eqn. 42 and approximate entropy eqn. 43 as follows,

$$\mathcal{G}^{\text{BETHE}}(\{q_i, \xi_{ij}\}) = E(\{q_i, \xi_{ij}\}) - S^{\text{BETHE}}(\{q_i, \xi_{ij}\}) \tag{46}$$

The expression for the entropy eqn. 43 is exact when the graph is a tree[4]. Since the expression for the energy eqn. 42 is exact for general Boltzmann machines this implies that the Bethe free energy is also exact on trees. For trees the probability distribution can be written as a function of the node and pairwise marginals only,

$$P(\{h_i\}) = \prod_{(ij)} p_{ij}(h_i, h_j) \prod_i p_i(h_i)^{1-z_i} \tag{47}$$

Computing the free energy of this probability distribution gives back eqn. 46. Expression 47 is of course not valid for graphs with loops, indeed it is not even a properly normalized probability distribution in that case, which implies that the Bethe free energy is not necessarily an upper bound to the exact free energy as in the MF case. So, when can we expect the Bethe free energy

[4] However, it may also become *negative* on a highly connected, highly correlated graph. For instance, 4 nodes with mean $\frac{1}{2}$ which are all connected and perfectly dependent, have -2 bits of entropy according to the Bethe approximation, while having $+1$ bit of entropy in reality.

to be a good approximation? The above argument suggests that this should be the case when the graph is "close to a tree", i.e. if there are not many short loops in the graph. In the case of tight loops, evidence impinging on one node can travel around these loops and return back to the original node, causing it to be over-counted. We will see another argument supporting this in section 8. There it will also be proved that the MF and Onsager approximations are small weight expansions of the Bethe approximation, suggesting that the Bethe approximation should be accurate for small weights and improve MF and Onsager. Intuitively, when the weights are small the evidence will not run around in the loops, but "dies out" before it feeds back into its node of origin. We have also observed that large thresholds, which represent the external evidence, tend to break the dependencies between neighboring nodes, and therefore improve the approximation. Summarizing, we could say that the approximation is good *when the correlation distance is shorter than the shortest loops in the system.* Small weights, large thresholds and long loops help achieve that.

On the other hand, in the limit of very large weights, the energy term will dominate the entropy term and the Bethe approximation should become exact. However, we have not observed good performance of either loopy belief propagation or belief optimization (see following sections) in this regime, possibly due to the fact that the energy surface becomes very complicated.

6.1 The Gibbs Free Energy in the Bethe Approximation

To make contact with the MF and Onsager approximations we will now convert the Bethe free energy eqn. 46 into an approximate Gibbs free energy. This is done by minimizing the Bethe free energy with respect to the correlations $\{\xi_{ij}\}$ and solving them exactly in terms of the marginals $\{q_i\}$. Taking derivatives of the Bethe free energy with respect to $\{\xi_{ij}\}$ and setting them to zero we find,

$$\frac{\partial \mathcal{G}^{\text{BETHE}}}{\partial \xi_{ij}} = -W_{ij} + \log\left(\frac{\xi_{ij}(\xi_{ij} + 1 - q_i - q_j)}{(q_i - \xi_{ij})(q_j - \xi_{ij})}\right) = 0 \tag{48}$$

This can be simplified to a quadratic equation,

$$\alpha_{ij}\xi_{ij}^2 - (1 + \alpha_{ij}q_i + \alpha_{ij}q_j)\xi_{ij} + (1 + \alpha_{ij})q_iq_j = 0 \tag{49}$$

where we have defined

$$\alpha_{ij} = e^{W_{ij}} - 1 \tag{50}$$

In addition to this equation we have to make sure that ξ_{ij} satisfies the following bounds,

$$\max(0, q_i + q_j - 1) \leq \xi_{ij} \leq \min(q_i, q_j) \tag{51}$$

These bounds can be understood by noting that eqn. 7,8,9 and 10, can not become negative. The following theorem provides the desired unique solution for $\{\xi_{ij}\}$.

Theorem 1 *There is exactly one solution to the quadratic equation (49) which satisfies the bounds (51). The analytic expression is given by*[5]*,*

$$\begin{aligned}
\xi_{ij} &= \frac{1}{2\alpha_{ij}}\left(Q_{ij} - \sqrt{Q_{ij}^2 - 4\alpha_{ij}(1+\alpha_{ij})q_i q_j}\right) \\
Q_{ij} &= 1 + \alpha_{ij} q_i + \alpha_{ij} q_j
\end{aligned} \tag{52}$$

Moreover, ξ_{ij} will never actually saturate one of the bounds.

Note that for $\alpha_{ij} \to 0$ we have $\xi_{ij} = q_i q_j$ which is the correct limit.

Proof of theorem 1

We will first prove that there must be exactly one minimum inside the bounds (51) (i.e. not on the bounds).

First, we compute the second derivative with respect to ξ_{ij},

$$\begin{aligned}
\frac{\partial^2 \mathcal{G}^{\text{BETHE}}}{\partial \xi_{ij}^2} &= \frac{1}{\xi_{ij}} + \frac{1}{\xi_{ij} + 1 - q_i - q_j} + \frac{1}{q_i - \xi_{ij}} + \frac{1}{q_j - \xi_{ij}} \\
&= \frac{1}{p_{ij}(1,1)} + \frac{1}{p_{ij}(0,0)} + \frac{1}{p_{ij}(1,0)} + \frac{1}{p_{ij}(0,1)} \geq 0
\end{aligned}$$

[5] For computational reasons it is sometimes convenient to use the following equivalent expression,

$$\xi_{ij} = \begin{cases} q_i q_j & \text{if } \alpha_{ij} = 0 \\ \frac{1}{2}\left(R_{ij} - \text{sign}(\beta_{ij})\sqrt{R_{ij}^2 - 4(1+\beta_{ij})q_i q_j}\right) & \text{otherwise} \end{cases}$$

where $\quad R_{ij} = \beta_{ij} + q_i + q_j \quad$ and $\quad \beta_{ij} = \frac{1}{\alpha_{ij}}$

Also from $\frac{\partial \mathcal{G}^{\mathrm{BETHE}}}{\partial \xi_{ij}}$ in eqn. 48 we see that at the lower boundary the derivative is $-\infty$ while at the upper boundary it is $+\infty$. Since the second derivative is always positive between the bounds and since the free energy is continuous between the bounds we infer that the free energy has exactly one minimum inside the bounds.

Next we proof that the positive root,

$$\begin{aligned} \zeta_{ij} &= \frac{1}{2\alpha_{ij}} \left(Q_{ij} + \sqrt{Q_{ij}^2 - 4\alpha_{ij}(1+\alpha_{ij})q_i q_j} \right) \\ Q_{ij} &= 1 + \alpha_{ij} q_i + \alpha_{ij} q_j \end{aligned} \tag{53}$$

to the quadratic eqn.49 is always located outside the bounds (except when $\alpha_{ij} = 0$ in which case the equation is degenerate). We can assume without loss of generality that $q_i \geq q_j$.

For $\alpha_{ij} = 0$ we have that the quadratic equation reduces to,

$$-\xi_{ij} + q_i q_j = 0 \tag{54}$$

with the obvious solution located within the bounds of eqn. 51. For $\alpha_{ij} > 0$ we will use the fact that,

$$\begin{aligned} & Q_{ij}^2 - 4\alpha_{ij}(1+\alpha_{ij})q_i q_j \\ = \;& 1 + 2\alpha_{ij} q_i (1-q_j) + 2\alpha_{ij} q_j (1-q_i) + \alpha_{ij}^2 (q_i - q_j)^2 \\ \geq \;& 1 + 2\alpha_{ij} \left(q_i(1-q_j) + q_j(1-q_i) \right) \\ \geq \;& 0 \end{aligned}$$

(this result is actually valid for all possible α_{ij}, i.e. in the range $(-1, \infty)$) The above result can now be used to prove,

$$\begin{aligned} \zeta_{ij} &\geq \frac{1}{2\alpha_{ij}} (1 + \alpha_{ij} q_i + \alpha_{ij} q_j) \\ &\geq \frac{1}{2\alpha_{ij}} + q_j \\ &> q_j \end{aligned} \tag{55}$$

which is always larger than the upper bound. Finally, for $\alpha_{ij} \in (-1, 0)$, we will use the fact that,

$$Q_{ij}^2 - 4\alpha_{ij}(1+\alpha_{ij})q_i q_j \geq Q_{ij} \tag{56}$$

with Q_{ij} defined in (53). This can be used to prove,

$$\begin{aligned}\zeta_{ij} &\leq \frac{1}{\alpha_{ij}} + q_i + q_j \\ &< -1 + q_i + q_j\end{aligned}$$

which is always smaller than the lower bound.

Therefore, since we know one of the solutions must be located at the minimum between the boundaries, and the positive root is always located outside the boundaries, we have proven that the negative root is precisely the valid solution, located at the minimum of the free energy, inside the boundaries. □

Thus, by substituting this expression for $\{\xi_{ij}\}$ back into the Bethe free energy we obtain the Gibbs free energy in the Bethe approximation:

$$G^{\text{BETHE}}(\{q_i\}) = \mathcal{G}^{\text{BETHE}}(\{q_i, \xi_{ij}(q)\}) \tag{57}$$

6.2 Belief Optimization

In the previous section we have derived an expression for the Gibbs free energy in the Bethe approximation, similar in spirit to the MF and Onsager expressions for the Gibbs free energy. We will now proceed to derive fixed point equations to solve the marginals $\{q_i\}$, which turn out to be generalizations of the MF and Onsager fixed point equations. We will call any algorithm that minimizes the Bethe free energy in primal space (i.e. in terms of the posterior probability distributions) *belief optimization* (BO) in the following.

We follow the by now familiar recipe: first compute derivatives of the Bethe-Gibbs free energy with respect to $\{q_i\}$ and then equate them to zero.

$$\frac{dG^{\text{BETHE}}}{dq_i} = \frac{\partial \mathcal{G}^{\text{BETHE}}}{\partial q_i} + \sum_{j\in N(i)} \frac{\partial \mathcal{G}^{\text{BETHE}}}{\partial \xi_{ij}} \frac{\partial \xi_{ij}}{\partial q_i} = \frac{\partial \mathcal{G}^{\text{BETHE}}}{\partial q_i} \tag{58}$$

since $\frac{\partial \mathcal{G}^{\text{BETHE}}}{\partial \xi_{ij}} = 0$ when the $\{\xi_{ij}\}$ are solved in terms of the $\{q_i\}$. Using the above we find,

$$\frac{\partial G^{\text{BETHE}}}{\partial q_i} = -\theta_i + \log\left(\frac{(1-q_i)^{z_i-1}\prod_{j\in N(i)}(q_i - \xi_{ij})}{q_i^{z_i-1}\prod_{j\in N(i)}(\xi_{ij} + 1 - q_i - q_j)}\right) \tag{59}$$

Equating these derivatives to zero gives the following set of fixed point equations,

$$q_i = \sigma\left(\theta_i + \sum_{j\in N(i)} \log\left(\frac{q_i(\xi_{ij}+1-q_i-q_j)}{(1-q_i)(q_i-\xi_{ij})}\right)\right) \tag{60}$$

These equations are direct generalizations of the MF and Onsager fixed point equations. When the expression inside the sigmoid is expanded for small weights and all terms up to quadratic order in the weights are retained we find the Onsager fixed point equations. When only terms up to first order are retained we find the MF equations.

Whether run sequentially or in parallel, the above equations are not guaranteed to decrease the Gibbs free energy or to converge at all. In analogy with the MF and Onsager equations we may achieve this by temporarily fixing all neighboring marginals q_j and minimizing over the central node q_i. We can show again that this subproblem is a convex minimization problem[6] and can be solved by many standard techniques. By cycling through all nodes we perform coordinate descent on the Bethe-Gibbs free energy. This idea is readily extended to an algorithm where each sub-problem is a convex minimization problem on a tree and a mix of BP and iterative scaling is used to solve it. In fact, this algorithm is applicable to general discrete networks and we refer to [25] for a more detailed description of that idea.

Alternatively, one can choose again to perform gradient descent on all the means simultaneously while enforcing the constraint that they stay within the interval $[0, 1]$.

6.3 The Bethe Approximation - Linear Response

To compute the covariances for neighboring nodes we can use $\xi_{ij} - q_i q_j$ where ξ_{ij} and q_i minimize the Bethe free energy. However, for non-neighbors (where $W_{ij} = 0$), this expression vanishes since $\xi_{ij} = q_i q_j$ as $W_{ij} = 0$. To improve the covariance estimates for both neighbors and non-neighbors we can use the linear response estimate eqn.24 again, applied to the Bethe approximation. Eqn.59 directly relates $\{q_i\}$ and $\{\theta_j\}$ at the minimum of the Bethe-Gibbs

[6] Since the central node i and its links to its neighbors is a tree the expression for the Bethe-Gibbs free energy of that region is exact. Moreover, the Gibbs free energy is the minimum of the variational free energy $F(Q)$, with linear constraints fixing the means to $\{q_i\}$. Since the variational free energy is convex in Q, the Gibbs free energy must be convex in $\{q_i\}$.

free energy. Taking derivatives of that expression with respect to q_i provides an expression for $\partial\theta_j/\partial q_i$ which is inverted to compute an approximation for the covariances,

$$[C_{ij}] = \left[\left(\frac{1-z_j}{q_j(1-q_j)} + \sum_{k\in N(j)} \left(\frac{1}{q_j-\xi_{jk}} + \frac{1}{\xi_{jk}+1-q_j-q_k} \right) \left(1 - \frac{\partial \xi_{jk}}{\partial q_j} \right) \right) \delta_{ji} \right.$$
$$\left. + \left(\frac{1}{\xi_{ji}+1-q_j-q_i} - \left(\frac{1}{q_j-\xi_{ji}} + \frac{1}{\xi_{ji}+1-q_j-q_i} \right) \frac{\partial \xi_{ji}}{\partial q_i} \right) \delta_{j,N(i)} \right]^{-1} \tag{61}$$

where $\delta_{j,N(i)}$ is 1 if i and j are neighbors and vanishes otherwise, and

$$\frac{\partial \xi_{ji}}{\partial q_j} = \frac{\partial \xi_{ij}}{\partial q_j} = \frac{1}{2}\left(1 - \frac{1+\alpha_{ij}(q_j-q_i)-2q_i}{\sqrt{Q_{ij}^2 - 4\alpha_{ij}(1+\alpha_{ij})q_iq_j}} \right) \tag{62}$$
$$Q_{ij} = 1 + \alpha_{ij}q_i + \alpha_{ij}q_j \tag{63}$$
$$\alpha_{ij} = e^{W_{ij}} - 1 \tag{64}$$

while ξ_{ij} is given by theorem 1. One can now check[7] that the Taylor expansion of this expression up to second order in the weights W_{ij} is identical to the Onsager result eqn.40. In fact, the Bethe approximation is not just a higher order truncation of the Plefka expansion, but contains contributions from all orders in W_{ij} (see section 8). Further, since the Bethe approximation is exact on trees for arbitrary thresholds $\{\theta_i\}$, the following statement must hold,

Theorem 2 *The covariances between two arbitrary nodes (including non-neighbors) on a Boltzmann tree, are exactly given by the linear response expression 61.*

This theorem is interesting because belief propagation will produce the exact covariances on trees only for neighboring nodes. The fact that *all* covariances are given by an analytic expression is somewhat surprising.

6.4 A Different Perspective

Much of the results presented in the previous subsections revolve around two equations, 48, relating the weights to the posterior marginals, and 59

[7] We used MAPLE to Taylor expand the expression for C_{ij}.

relating the thresholds to the posterior marginals. It is well known that the Boltzmann machine, being an exponential model, can be parameterized with either the weights and thresholds or with the sufficient statistics $\langle h_i \rangle = q_i$ and $\langle h_i h_j \rangle = \xi_{ij}$ [1]. The above mentioned equations relate the parameters and the sufficient statistics in the Bethe approximation. Therefore all we have been doing is to invert these equations in order to compute the sufficient statistics in terms of the parameters.

Could we have predicted eqns. 48 and 59 without making reference to the Bethe free energy? With hindsight, yes. First, let's try to change parameterization on a Boltzmann *tree*. For every pair of nodes we reparameterize as follows,

$$p_{ij}(h_i, h_j) = \exp(K_{ij} h_i h_j + H_{ji} h_i + H_{ij} h_j + C_{ij}) \tag{65}$$

This can then be inverted,

$$\begin{aligned}
K_{ij} &= \log\left(\frac{p_{ij}(h_i = 1, h_j = 1) p_{ij}(h_i = 0, h_j = 0)}{p_{ij}(h_i = 1, h_j = 0) p_{ij}(h_i = 0, h_j = 1)}\right) && (66)\\
H_{ji} &= \log\left(\frac{p_{ij}(h_i = 1, h_j = 0)}{p_{ij}(h_i = 0, h_j = 0)}\right) \qquad H_{ij} = \log\left(\frac{p_{ij}(h_i = 0, h_j = 1)}{p_{ij}(h_i = 0, h_j = 0)}\right) && (67)\\
C_{ij} &= \log\left(p_{ij}(h_i = 0, h_j = 0)\right) && (68)
\end{aligned}$$

Analogously, for a node marginal we write,

$$p_i(h_i) = \exp(H_i h_i + B_i) \tag{69}$$

which can again be solved to give,

$$\begin{aligned}
H_i &= \log\left(\frac{p_i(h_i = 1)}{p_i(h_i = 0)}\right) && (70)\\
B_i &= \log\left(p_i(h_i = 0)\right) && (71)
\end{aligned}$$

But on a tree we know the form of the full joint distribution in terms of the node and pairwise marginals, namely eqn. 47. This should be equal to

$$P(\{h_i\}) = \exp(\sum_{(ij)} W_{ij} h_i h_j + \sum_i \theta_i h_i + F) \qquad F = -\log(Z) \tag{72}$$

Combining all this gives the following relations,

$$W_{ij} = K_{ij} \tag{73}$$

$$\theta_i = \sum_{j \in N(i)} H_{ji} + (1 - z_i) H_i \tag{74}$$

$$F = \sum_{(ij)} C_{ij} + \sum_i (1 - z_i) B_i \tag{75}$$

Finally inserting the expressions 66, 67, 68, 70 and 71 into the above expression, and using the definitions of section 2, we precisely arrive at 48 and 59. Although the conversion from marginal posterior probabilities to weights and thresholds is very simple, the transformation the other way round is obviously not so easy, since it is precisely solved by belief propagation and belief optimization. In the Bethe *approximation*, these same conversion rules are now employed on graphs with loops.

7 Loopy Belief Propagation

As an alternative to the above procedure we give a simple derivation of a set of fixed point equations equivalent to the equations for loopy belief propagation (BP). We follow the derivation in [14].

We shall begin with writing expressions for the node and pairwise marginals in terms of the weights, thresholds and "mean fields" μ_{ij} that parameterize the influence of neighboring nodes.

$$p_i(h_i) \propto e^{\left(\theta_i h_i + \sum_{k \in N(i)} \mu_{ki} h_i\right)} \tag{76}$$

$$p_{ij}(h_i, h_j) \propto e^{\left(W_{ij} h_i h_j + \theta_i h_i + \theta_j h_j + \sum_{k \in \{N(i) \setminus j\}} \mu_{ki} h_i + \sum_{l \in \{N(j) \setminus i\}} \mu_{lj} h_j\right)} \tag{77}$$

Next, we solve for the mean fields by requiring that,

$$\sum_{h_j} p_{ij}(h_i, h_j) = p_i(h_i) \tag{78}$$

for all i and j. Inserting eqns. 76 and 77 into the consistency equation 78, we arrive at the following fixed point equations,

$$\mu_{ji} = \log\left(1 + \alpha_{ij}\ \sigma\left(\theta_j + \sum_{l \in N(j)} \mu_{lj}\right)\right) \tag{79}$$

where $\alpha_{ij} = e^{W_{ij}} - 1$ was defined in 50.

The mean fields μ turn out to be the logarithm of the messages in belief propagation as defined in [32]. In that paper it was elegantly shown *that the fixed points of the belief propagation updates, equivalent to 79, correspond to the stationary points of the Bethe free energy.* Lagrange multipliers were used to enforce the consistency constraints 78 and the mean fields (or messages) turn out to be simple functions of those Lagrange multipliers.

An important property of the belief propagation updates is that on trees they converge within a finite number of iterations (equal to the length of the longest path in the tree) to the exact result.

8 Plefka's Expansion Revisited

In the previous sections we have encountered 3 approximations, the Mean Field, Onsager and Bethe approximations. In this section we hope to further clarify their relationship through the Plefka expansion of the Gibbs free energy which was explained in section 4. There we have seen that the Onsager approximation contains precisely all terms up to second order in the weights, while the mean field approximation contains all terms up to first order. The question to be answered is if we can characterize the Bethe approximation using this Plefka expansion. In this section we will prove George's and Yedidia's conjecture, [3],[31] which states that the Bethe approximation consists of an infinite subset of terms in the Plefka expansion, namely exactly those which contain one or two nodes. We start with the proof of a more general result, also conjectured by Georges and Yedidia, stating that the Plefka expansion only contains so called "strongly irreducible diagrams". By a diagram we mean a graph drawn for a particular term in the Plefka expansion where the vertices correspond to the nodes present in that term and edges correspond to weights W_{ij} between nodes (a separate edge is drawn for every W_{ij} in the term). Strongly irreducible diagrams corresponds to graphs that have the property that removing any node from the graph does not split it in two pieces. As an example of a strongly irreducible diagram we have drawn Onsager's reaction term in figure 1a.

Theorem 3 *Plefka's expansion of the exact Gibbs free energy has no strongly reducible diagrams.*

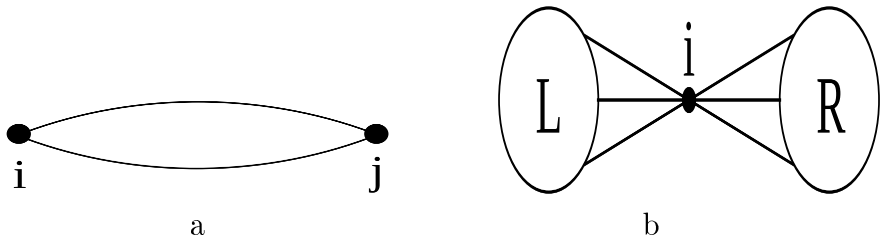


Figure 1: Diagram for Onsager's reaction term (a) and hypothetical bottleneck diagram (b).

Proof of Theorem 3
First let us assume that there exists a strongly reducible diagram in the Plefka expansion. Then, construct a subgraph of the original graph by setting all weights that do not appear in the diagram to zero, and removing the disconnected nodes. Construct the Plefka expansion of the subgraph by setting to zero these same weights in the Plefka expansion of the full graph, and removing the disconnected single node terms. In the above construction, the strongly reducible diagram will appear in the expansion of the subgraph and has the general form depicted in figure 1b, which we named a bottleneck-diagram where the bottleneck node i is the node which is causing the diagram to be strongly reducible (i.e. removing it cuts the diagram in two).

From the structure of the subgraph we can infer that its probability distribution can be written as follows,

$$P(h_L, h_i, h_R) = P_L(h_L|h_i)\ P_R(h_R|h_i)\ P_i(h_i) \tag{80}$$

implying that the energy can be written as[8],

$$E(\{h_j\}) = E_L(h_L, h_i) + E_R(h_R, h_i) + E_i(h_i) \tag{81}$$

We will now show that the variational formulation of the Gibbs free energy together with the above decomposition of the energy and the constraint that the marginal of node i is fixed, are enough to show that the Gibbs free energy decomposes as in eqn. 88. Recall that the Gibbs free energy is defined as,

$$G(\{\hat{Q}_j\}) = \min_Q \left\{ \langle E \rangle_Q - S(Q) + \sum_j \sum_{h_j} \lambda_j(h_j)(Q(h_j) - \hat{Q}_j(h_j)) \right\} \tag{82}$$

[8] Actually for BMs the energy decomposed even further as $E = \sum_{(ij)} E_{(ij)}$.

where the Lagrange multipliers $\{\lambda_j(h_j)\}$ should be chosen so that they enforce the marginalization constraints. Equating the functional derivatives of G with respect to Q to zero and using the energy decomposition eqn.81, we arrive at the following form for Q,

$$Q(\{h_j\}) = \frac{1}{Z} e^{-\tilde{E}_L(h_L,h_i)}\ e^{-\tilde{E}_R(h_R,h_i)}\ e^{-\tilde{E}_i(h_i)} \tag{83}$$

where Z is the normalization constant and we have defined,

$$\tilde{E}_L(h_L,h_i) = E_L(h_L,h_i) + \sum_{l\in L} \lambda_l(h_l) \tag{84}$$

$$\tilde{E}_R(h_R,h_i) = E_R(h_R,h_i) + \sum_{l\in R} \lambda_l(h_l) \tag{85}$$

$$\tilde{E}_i(h_i) = E_i(h_i) + \lambda_i(h_i) \tag{86}$$

In other words, the presence of the Lagrange multipliers hasn't changed the form of the energy decomposition eqn.81, which means that Q factorizes in precisely the same way as P in eqn.80. The constraint that the bottleneck node i has a marginal $\hat{Q}_i$ is trivially imposed leading to the following form for Q,

$$Q(\{h_j\}) = Q_L(h_L|h_i)\ Q_R(h_R|h_i)\ \hat{Q}_i(h_i) \tag{87}$$

Feeding this expression for Q back into eqn.82 and using once again the energy decomposition eqn.81 we find that the Gibbs free energy decomposes as

$$\begin{aligned} G(\{\hat{Q}_j\}) &= \min_{Q_L}\left\{ \langle E_L + \log(Q_L)\rangle_{Q_L\hat{Q}_i} + \sum_{j\in L}\sum_{h_j} \lambda_j(h_j)(Q_j(h_j) - \hat{Q}_j(h_j)) \right\} \\ &+ \min_{Q_R}\left\{ \langle E_R + \log(Q_R)\rangle_{Q_R\hat{Q}_i} + \sum_{j\in R}\sum_{h_j} \lambda_j(h_j)(Q_j(h_j) - \hat{Q}_j(h_j)) \right\} \\ &+ \langle E_i + \log(\hat{Q}_i)\rangle_{\hat{Q}_i} \\ &= G_L + G_R + G_i \end{aligned} \tag{88}$$

where $Q_L\hat{Q}_i = Q_L(h_L|h_i)\hat{Q}_i(h_i)$ and similarly for $Q_R\hat{Q}_i$. But, because the strongly reducible diagram is part of the Plefka expansion, and it contains nodes in both L and R, the Gibbs free energy can never decompose into the above form. This implies that we have proven a contradiction based on the assumption that the strongly reducible diagram is part of Plefka's expansion, which must therefore be false. □

Notice that the above decomposition of the Gibbs free energy does not imply that systems L and R are independent. What it does imply is that changes in the parameters of one subsystem do not influence the probabilities of the other subsystem. For instance, assume that the bottleneck node is the only node whose marginal is fixed. Changing the evidence impinging on some of the nodes of subsystem L will change the thresholds and subsequently all marginal posterior probabilities of that system. However, these changes do not affect the marginal posterior probabilities of the other subsystem, since they are the marginals of the distribution $Q_R \hat{Q}_i$ that minimizes G_R. The minimizations over Q_L and Q_R are independent since $Q_i = \hat{Q}_i$ is fixed. In other words, the Lagrange multiplier associated with the constrained bottleneck node counteracts the changes of the parameters so that its effect on the posterior probabilities remain isolated to the subsystem in which they occurred. The same effect was exploited in [25] to formulate a convergent coordinate descent algorithm (the UPS algorithm) on more general undirected graphical models.

We are now ready to formulate the precise relation between the Bethe free energy and the Plefka expansion,

Theorem 4 *The Plefka expansion of the Gibbs free energy in the Bethe approximation contains precisely all diagrams with one or two nodes present in the Plefka expansion of the exact Gibbs free energy.*

Proof of Theorem 4

We have already seen that the Bethe free energy is a sum of terms, where each term either depends on a marginal or a pairwise marginal. In addition, we have seen that the pairwise marginals can be solved in terms of their neighboring marginals only. Substituting these solutions into the Bethe free energy then shows that there are only terms in the Bethe-Gibbs free energy that depend on neighboring marginals.

Next, consider any two neighboring nodes. Construct a subgraph of the original graph by setting all weights, except the one connecting the two nodes in question to zero, and removing all other nodes. Set the same weights to zero in both the exact expansion as well as in the Bethe expansion (and remove disconnected single node terms). Since the Bethe approximation is exact on that diagram (it is a tree), their expansions must be equal. Applying this argument to any two nodes implies that all the diagrams consisting of one or two nodes in the exact expansion must also be present in the expansion of the Bethe-Gibbs free energy. $\square$

Finally, the relation between the MF, Onsager and Bethe approximations is clarified by the following theorem,

Theorem 5

$$G^{BETHE} = G^{ONSAGER} + O(W_{ij}^3) \tag{89}$$

$$G^{ONSAGER} = G^{MF} + O(W_{ij}^2) \tag{90}$$

This can be proven by expanding out the Bethe-Gibbs free energy up to second order in the weights and checking the result against the expressions for the Onsager- and MF-Gibbs free energies. Therefore, since the Bethe approximation includes more terms of the Plefka expansion than the Onsager approximation, which in turn includes more terms than the MF approximation, we expect that the accuracy of the approximations behaves accordingly, at least when the Plefka expansion exists (i.e. converges).

From the above result, it immediately follows that the fixed point equations to compute $\{q_i\}$ in the Bethe approximation reduce to the fixed point equations for the Onsager approximation when expanded to second order in the weights, which in turn reduces to MF when only linear terms are included. The analogous results for the linear response estimates of the covariances were already established.

One useful conclusion we can draw from the above theorems is the following,

Corollary 1 *If the set of links of the shortest loop in a graph is denoted as L, and the Plefka expansion converges, then the difference between the exact Gibbs free energy and the Bethe approximation of the Gibbs free energy is at least of the order $\prod_{(ij)\in L} W_{ij}$.*

9 Experiments

To assess the quality of the various approximations introduced in this paper we computed the mean activation q_i and covariance $\xi_{ij} - q_i q_j$ for all nodes and pairs of neighboring nodes on a 10×10 square lattice, where only nearest neighbors were connected. Weights and thresholds were sampled from a zero mean Gaussian distribution with varying standard deviations s_W and s_θ respectively. The thresholds were also shifted by an amount $\theta_i \to \theta_i - \frac{1}{2}\sum_{j \in N(i)} W_{ij}$, such that in a network with no external evidence, a node with zero (shifted) threshold will have a mean of $\frac{1}{2}$. Exact computations are still feasible on these graphs using the junction tree algorithm, where each row is converted into a super-node.

9.1 Comparing MF, Onsager, BP, and BO

In this experiment four methods were compared against the exact junction tree algorithm on a 10×10 lattice. The fixed point equations for MF (20), Onsager (39), BP (79) and BO (60) were used to compute the means while the linear response expressions for MF (25), Onsager (40) and BO (61) and expression 77 for BP were used to compute the covariances for neighboring nodes. We also computed the approximate free energies by inserting the means and correlations at convergence into the respective Gibbs free energies. For MF, Onsager and BO the code was halted when the maximum change in the means was smaller than 10^{-10} or 1000 iterations were performed. For BP analogous stopping criteria were used for the messages. For MF, Onsager and BO the means were randomly initialized between 0 and 1, while all messages were initialized at 1. To promote convergence we applied linear damping on the fixed point equations slowly increasing from 0 to 0.9. The absolute error averaged over 10 random draws of the networks for the means q_i (averaged over all nodes), neighboring covariances C_{ij} (averaged over all pairs of neighboring nodes) and free energy F are shown in figure 2 (standard deviation of thresholds is 1) and 3 (standard deviation of thresholds is 5). We also show the errors on a logarithmic scale in order to zoom in on the results for small weights. From these results we may conclude that in this regime of the weights, i.e. between 0 and 4, BP and BO converge to exactly the same means. Moreover, these BP/BO estimates are always better than Onsager and MF, while Onsager in turn is always more accurate than MF. One can also observe that the linear response estimates in the Bethe approximation of the covariances for neighboring nodes always improves the covariance estimates given by BP, Onsager and MF. When the weights grow, and the thresholds are kept constant, the results deteriorate. The same is true when the thresholds become smaller and the weights are kept constant

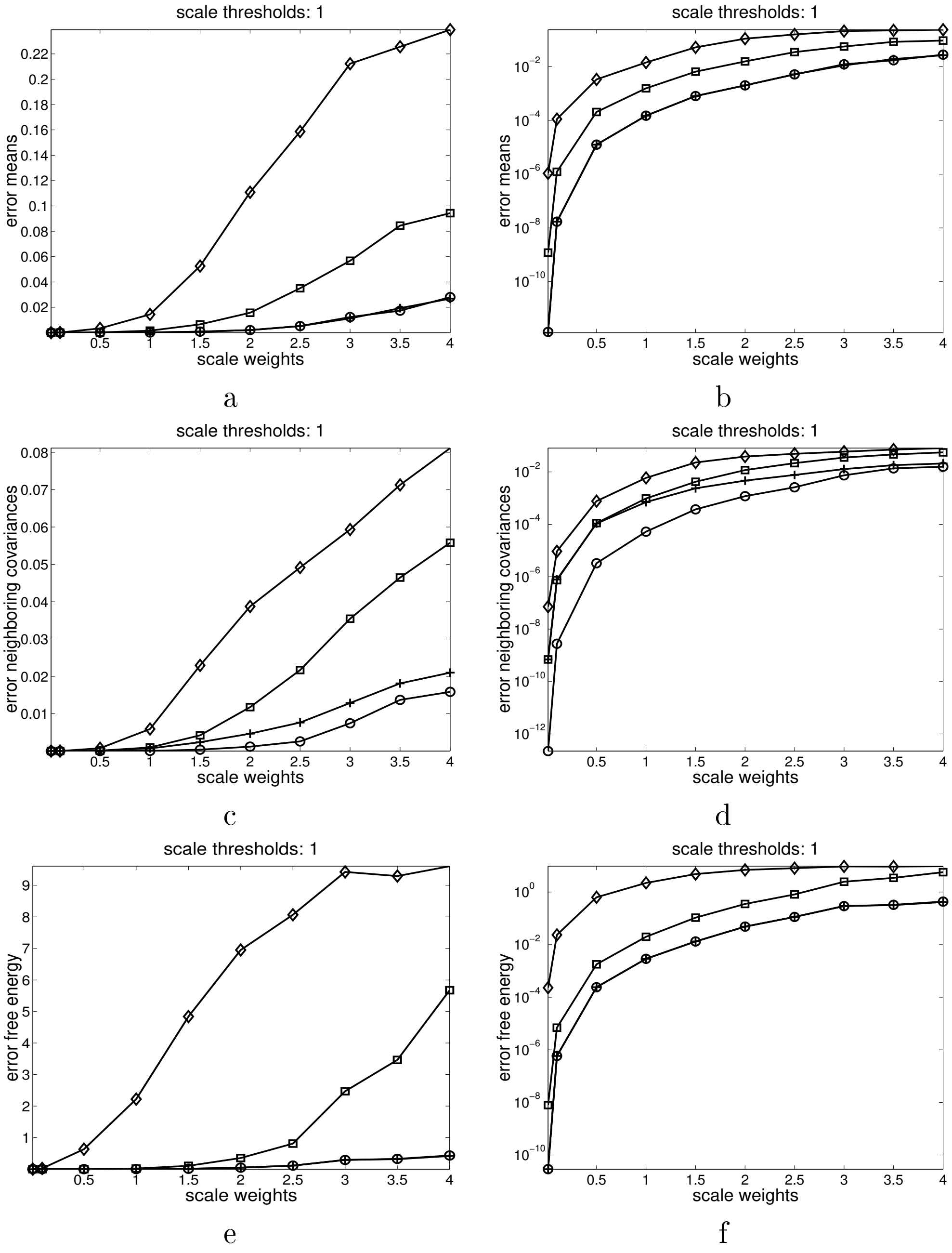


Figure 2: Absolute errors in means (a),(b), covariances for neighboring nodes (c),(d), and free energy (e),(f) on a 10×10 grid. Diamonds ("$\diamond$") indicate MF data, squares ("$\square$") Onsager data, crosses ("+") BP data and circles ("o") BO data. Right hand plots are on a log-scale. BP and BO results sometimes coincide.

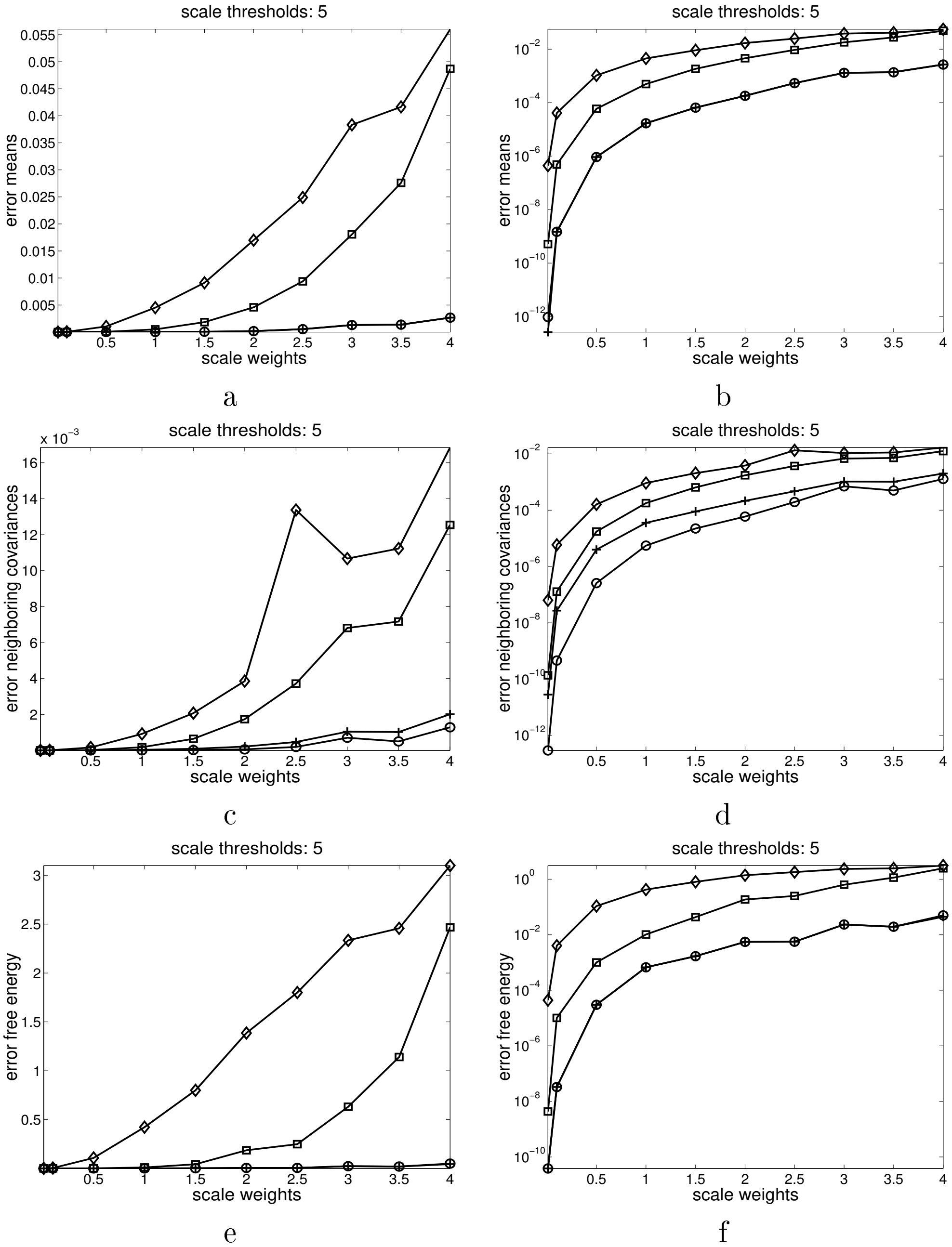
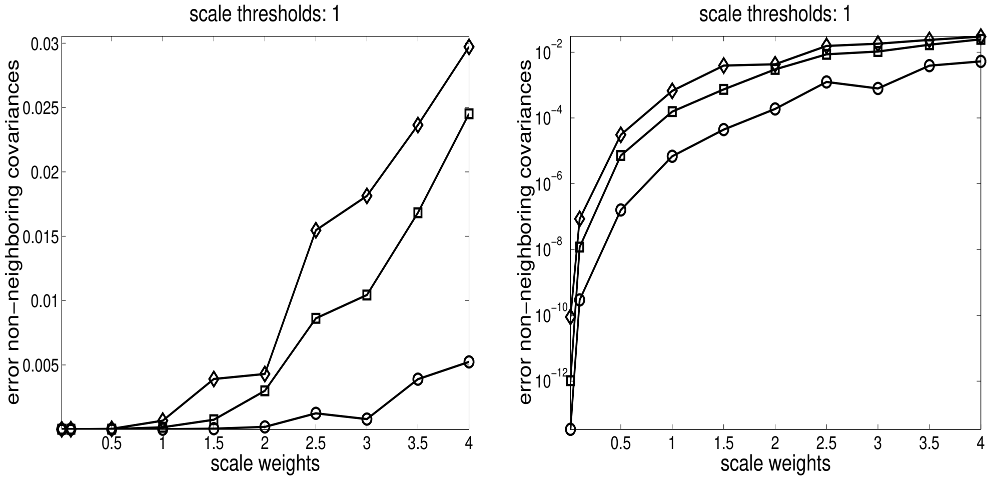


Figure 3: Same as in figure 2 with scale of thresholds fixed at 5.



a b

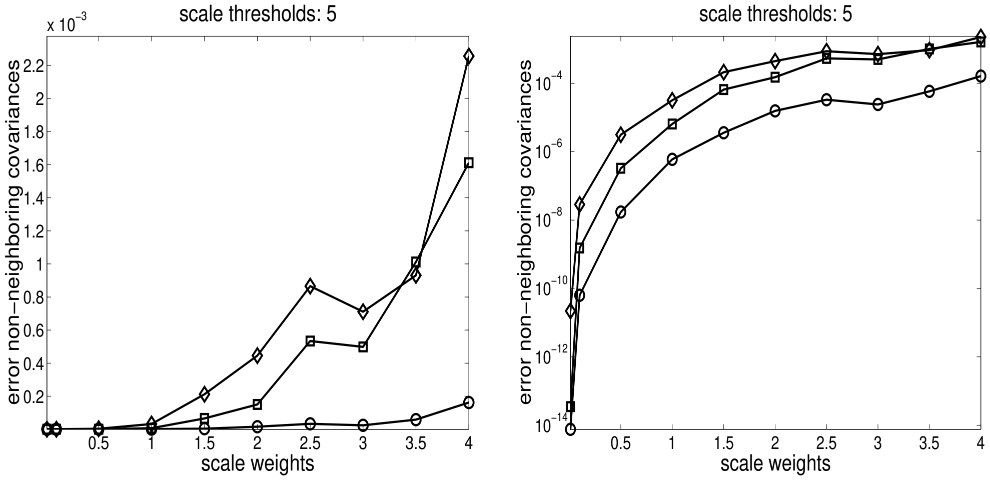
Figure 4: Linear (a) and logarithmic (b) plots of covariances for non-neighboring nodes on a 4×4 square lattice with scale of thresholds fixed at 1. Everything else is as in figure 2.

(note that the errors for $s_\theta = 5$ are much smaller than for $s_\theta = 1$). The regime where all methods fail to produce accurate estimates is when the weights are large and the thresholds small. In this case the energy surface is expected to be highly multi-modal. A more thorough analysis of the behavior of BP and BO in this regime is deferred to the next section.

To compare the covariances of non-neighbors we had to resort to smaller (4×4) networks, since they are not readily computed using the junction tree algorithm. Instead, exact values were computed by exhaustive summation over all 2^{16} states. The respective linear response estimates were computed for MF, Onsager and BO and compared with the exact computation. Figures 4 and 5 show these results for thresholds with scale 1 and 5 respectively. We conclude again that the linear response estimates of the Bethe approximation improve the linear response estimates of MF and Onsager considerably.

9.2 A Better Look at BP versus BO

In this section we compare the performance of BO and BP on a 10×10 square lattice. The standard deviations s_W and s_θ in this experiment were chosen from $\{0.1, 1, 3, 6, 10\}$ separately. For each setting of s_W and s_θ, 20



a b

Figure 5: Same as in figure 4 with scale of thresholds fixed at 5.

networks are generated to compare BP and BO. For large weights $s_W \geq 6$ and small thresholds $s_\theta \leq 3$, we generated and used 40 networks instead, as BP does not converge all the time.

For BO, we iterate over the nodes i of the network, clamping the neighboring marginals $q_j, j \in N(i)$ to their current values and running iterative scaling on the star-shaped segment to solve for the marginal q_i. This algorithm has the advantage that it is guaranteed to converge to a local minimum of the Bethe free energy. For BP, we used a strong damping factor of 0.9 so that it has a higher chance of convergence. For both algorithms, the convergence criteria was for all the means q_i to be changed by less than 10^{-4} for twenty consecutive iterations. We stop if BP has not converged by 10000 iterations.

For a given setting of s_W and s_θ, the generated networks are separated into two sets: one in which BP converged, and one in which it failed to converge. For each set separately, we compared BO and BP using the mean error in the estimated marginals q_i averaged over all nodes and all networks in the set. We also compared the mean error in the estimated covariances $\xi_{ij} - q_i q_j$, averaged over all neighboring pairs of nodes and all networks in each set[9]. Accompanying each mean we also looked at the mean absolute

[9]Note that the linear response correction was not used here.

deviation (MAD). The results are given in figure 6.

Each row of figure 6 corresponds to a setting of s_W, increasing from top to bottom. Within each row the left plot shows the errors in the estimated marginals, while the right plot shows the errors in the estimated covariances. In each plot there are five groups of four bars each. Each group corresponds to a setting of s_θ, increasing from left to right. In each group, the first two bars show the errors using BO and BP respectively, when BP converged. The next two bars show the errors for both BO and BP when BP failed to converge in 10000 iterations. The number associated with each group indicates the percentage of runs that BP failed to converge.

The qualitative behavior of the errors of the marginals and covariances are the same. Hence we shall concentrate on the errors of the marginals. The general trends in figure 6 confirm our expectations. With increasing weights, both BP and BO performed increasingly worse, as the distribution becomes more complicated and multi-modal. With increasing thresholds, both BP and BO performed better, as the distribution tends toward a single mode.

For small weights or large thresholds ($s_W \leq 3$, or $s_\theta \geq 6$, or $s_W = 6$ and $s_\theta \geq 1$), BP always converged and both algorithms performed equally well. As a matter of fact, most of the time both algorithms converged to very similar solutions. This is shown in the left plot of figure 7, where we plot the marginals obtained by BO versus those obtained by BP. In the right plot of figure 7, the algorithms sometimes get stuck in local minima or plateaus, resulting in a very small number of marginals being different : out of a total of 12000 points on the plot, only 372 lie outside the region $|x - y| \leq 0.01$.

The situation is more complicated for large weights and small thresholds ($s_W = 6$ and $s_\theta = 0.1$, or $s_W = 10$ and $s_\theta \leq 3$). In the regime where $s_\theta = 0.1$ and $s_W = 6, 10$, BO performed better than BP, especially when $s_W = 10$. In the regime where $s_\theta \geq 1$ and $s_W = 10$, BP amazingly performed better than BO even when BP did not converge.

One possible explanation for this phenomenon is that BO is stuck in local minima or plateaus, in which case we can diagnose this by seeing if the Bethe free energy of the final beliefs obtained using BO is larger than the Bethe free energy of the beliefs obtained using BP[10]. This is shown in figure 8. We see that the reverse is true instead – BO always converges to a point where

[10]When BP did not converge we computed the Bethe free energy by computing the means $\{q_i\}$ using eqn.76 and the correlations $\{\xi_{ij}\}$ using theorem 1 (instead of eqn.77) to ensure that they are consistent.

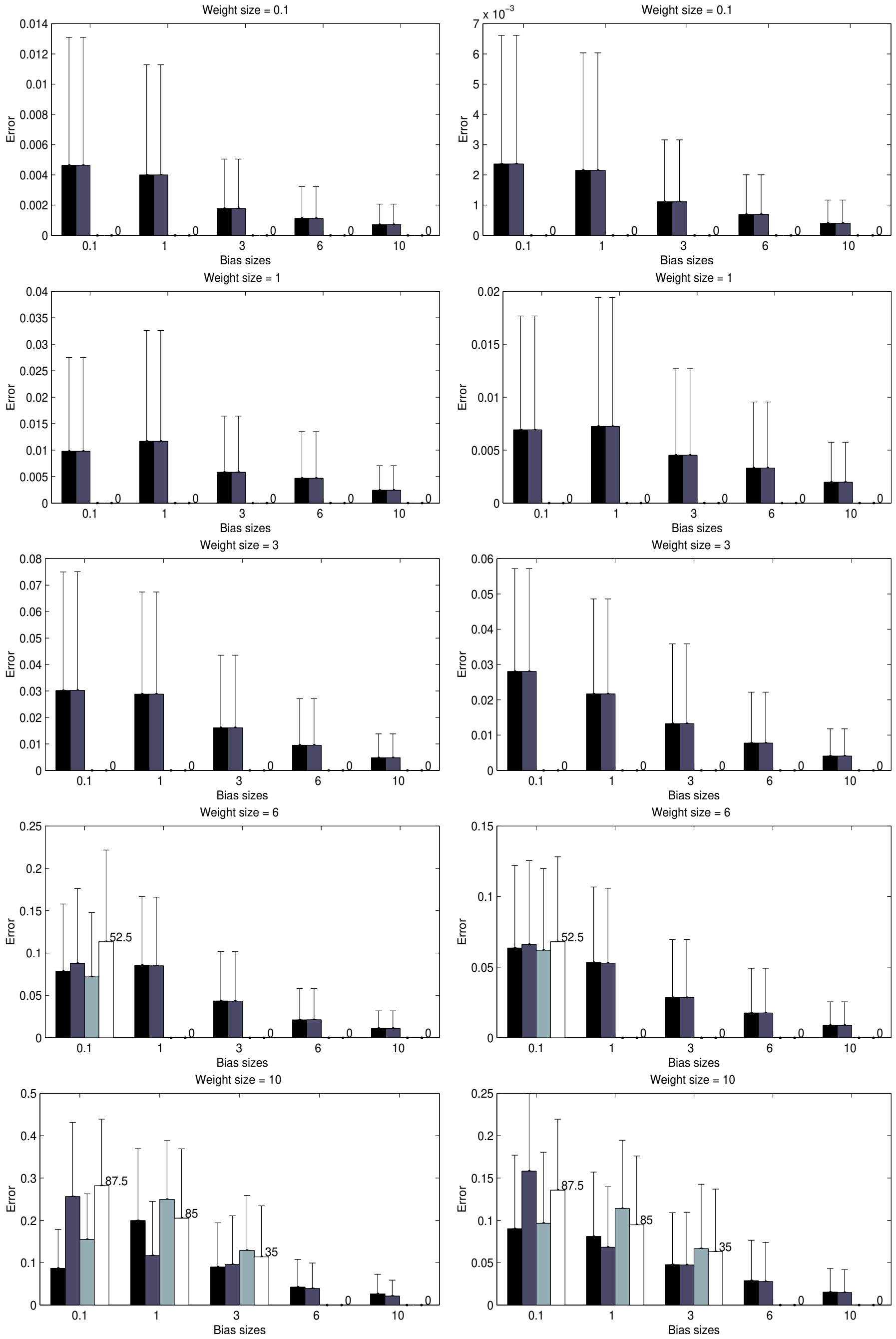


Figure 6: Errors in the estimated marginals (left) and the estimated covariances between neighboring nodes (right).

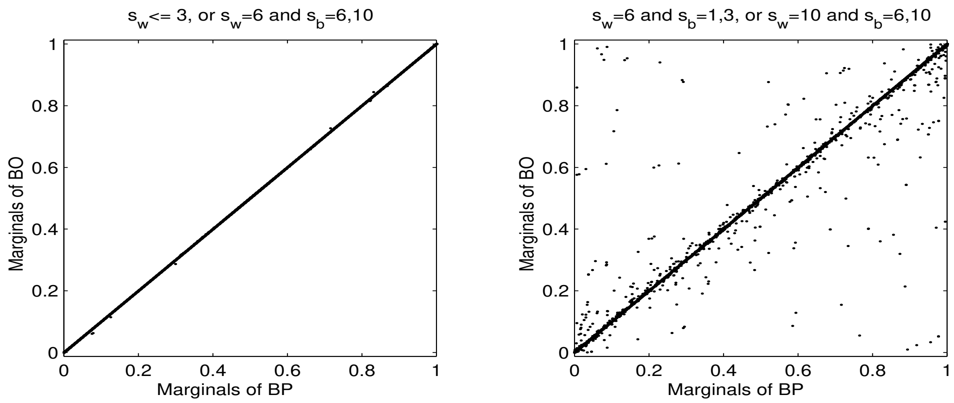


Figure 7: Scatter plots of the estimated marginals $b_i(1)$ for BP on the x-axis and for BO on the y-axis. The left plot is for networks with $s_W \leq 3$ or $s_W = 6$ and $s_\theta = 6, 10$. The right plot is for networks with $s_W = 6$ and $s_\theta = 1, 3$, or $s_W = 10$ and $s_\theta = 6, 10$.

the Bethe free energy is lower than the Bethe free energy obtained with BP. This shows that BO is not stuck in local minima and also shows that BO does what it was advertised to do – to decrease the Bethe free energy.

To understand why BO gives larger errors than BP we look at how the marginals estimated by BO and BP are related to the true marginals. This is shown in figure 9. We did not separate out cases where BP converged from those where it did not because the analyses turn out to be similar. Consider first the left plot of BP marginals versus the true marginals. Most of the points are concentrated near the $(1, 1)$ and $(0, 0)$ corner. This means that if a true marginal is close to 0 or 1, BP often converges to a limit cycle or stationary point close to the true marginal. Otherwise the BP marginal can be totally unrelated to the true marginal, as seen by the uniform spread of the points on the plot away from $(0, 0)$ and $(1, 1)$. In summary, BP often got the right marginal but sometimes got it totally wrong. Now consider the right plot of BO marginals versus the true marginals. Since there are not many points in the top left and bottom right quadrants, we see that the BO marginals are often on the same side of 0.5 as the true marginals. The problem lies with the (almost) horizontal ridge of points, where regardless of what the true marginal is, the BO estimate is often close to 0.5 (even though the BO estimate might lie on the same side of 0.5). This is true even when the true marginal is near to 0 or 1 (observe the two clumps of points, one

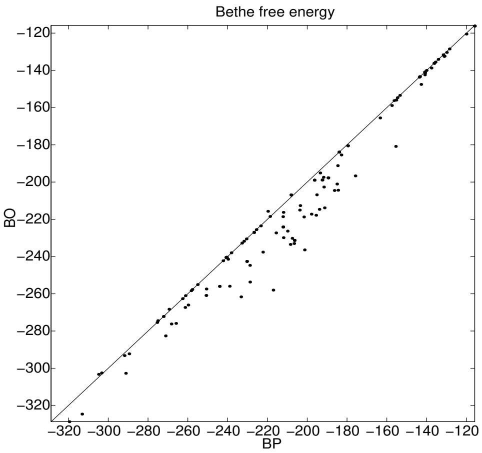


Figure 8: Scatter plot of Bethe free energy obtained using BO versus those obtained using BP.

near $(0, 0.5)$ and one near $(1, 0.5)$). It is the points near $(0, 0.5)$ and $(1, 0.5)$ which contributed to the high error as report in figure 6. BO prefers its marginals to be near 0.5 because they give a lower Bethe free energy, as seen in figure 8.

The same analysis shows why BO does better than BP when $s_W = 6, 10$ and $s_\theta = 0.1$. The results are shown in figure 10 for $s_W = 10$. The results are similar for $s_W = 6$. Again we did not distinguish between whether BP converged or not as the analyses were similar. First of all note that because the thresholds are so small, the true marginals are mostly between 0.3 and 0.7. Loopy BP did not converge in 87.5% of the networks, and we can see from figure 10 that the marginals it estimated are essentially random. For BO, the points in the right plot of figure 10 can be approximately split into two strips : a horizontal strip from $(0, 0.5)$ to $(1, 0.5)$, and a less distinct vertical strip from $(0.5, 0)$ to $(0.5, 1)$. This means that BO marginals are either close to 0.5 (horizontal strip), or are totally random (vertical strip). This should not be much better than what BP did on the left plot. The reason the BO errors in figure 6 are so much smaller than the BP errors is because the true marginals themselves are coincidentally often close to 0.5.

The above detailed analysis shows that BP always converges when the Bethe approximation is good and both BO and BP will converge to the same solution in this case. If however the Bethe approximation is bad, BP often does not converge, but BO does not seem to do much better either.

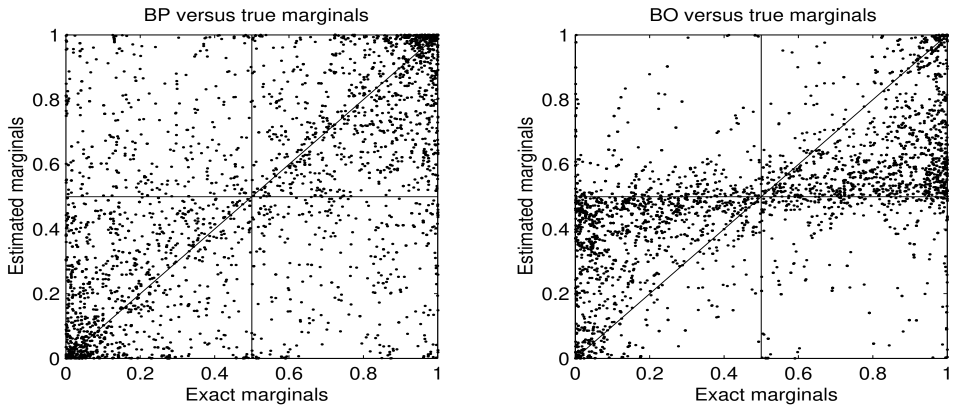


Figure 9: Scatter plot of BP and BO marginals versus the true marginals for networks with $s_W = 10, s_\theta = 1$.

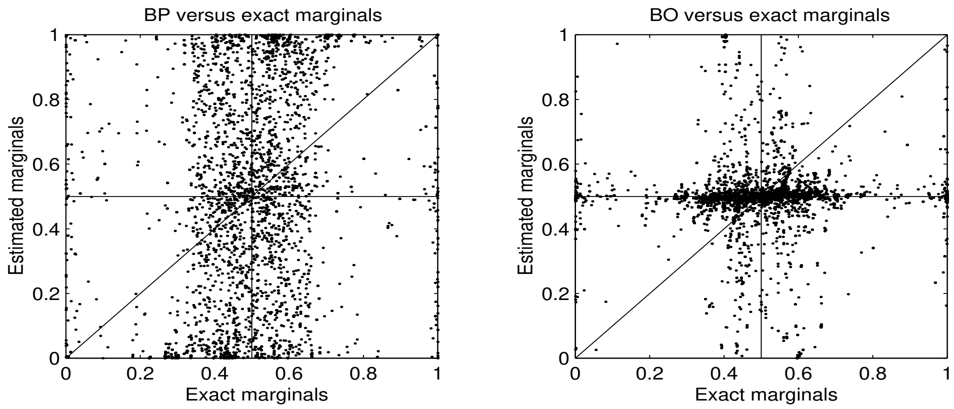


Figure 10: Scatter plot of BP and BO marginals versus the true marginals for networks with $s_W = 10, s_\theta = 0.1$.

10 Discussion

In this paper we have reviewed the mean field, Onsager and Bethe approximations. Our contribution was to convert the Bethe free energy to a Gibbs free energy by solving the pairwise marginals ξ_{ij} in terms of the neighboring marginals q_i and q_j. This resulted in a new algorithm to minimize the Bethe free energy in "primal" space (i.e. directly in terms of the node and pairwise marginals), which forms a direct generalization of the fixed point equations for the MF and Onsager approximations. Moreover, provably convergent algorithms were derived to directly minimize the Bethe free energy. A further result of this primal formulation allowed us to improve the estimates of the correlations through the use of the linear response theorem. Finally we proved a number of long standing conjectures concerning the Plefka expansion of spin systems (i.e. the Boltzmann machine), which further clarified the relationship between the MF, Onsager and Bethe approximations.

A notable difference between BP and BO is the fact that BP need not satisfy the marginalization constraints before it has converged. In contrast, BO is parameterized such that it will satisfy these constraints automatically. The above implies that the dynamics by which BP and BO try to minimize the Bethe free energy are of a very different nature. An undesirable property of BP, namely its failure to converge under certain circumstances, is certainly avoided by BO. However, the general conclusion from our experiments is that the Bethe approximation probably breaks down before any significant difference between the two methods shows up.

In previous work we have also developed BO algorithms for the Gaussian case and the non-binary discrete case [29],[25]. For Gaussian belief propagation (GaBP) it is important to notice that message updates do not necessarily maintain positive definiteness of the covariance matrix. This does not come as a surprise since it is a global constraint, while BP only performs local computations. As a consequence, the Bethe free energy is *not* always bounded from below and we have observed that exactly in these cases both GaBP and Gaussian BO (GaBO) do not converge. In all other cases GaBP and GaBO find the same answer experimentally. For a certain class of interactions (diagonally dominant) it was proved in [28] that GaBP always converges. An algorithm for undirected graphical models with more than two states per node, named "Unified Propagation and Scaling" (UPS) was proposed in [25]. It segments the graph in a forest of trees by fixing the marginals of certain nodes. A combination of iterative scaling and BP is then performed on these trees. Next a new set of nodes is clamped to their current marginal posterior estimates, resulting in a new forest of trees, etc. When no nodes remain frozen all the time, the UPS algorithm will converge to a stationary point of the Bethe free energy. An interesting alternative is Yuille's CCCP algorithm [33], which is also guaranteed to converge to a stationary point of the Bethe free energy.

The Bethe approximation is actually the simplest of a whole family of "Kikuchi" approximations which treat larger clusters exactly. In [32] the "generalized belief propagation" algorithm was derived to find stationary points of these Kikuchi approximations. Extending the ideas presented in this paper to these more accurate approximations is a topic of future research.

Another direction for future research is the extension of the theorems proven in section 8 to more general situations. One evident consequence of these theorems is the following. Consider a graph which has a bottleneck of more than one node, and moreover the joint distribution of these nodes is kept fixed (not just the node marginals of these nodes). Then, changes in the parameters of a subgraph on one side of the bottleneck (e.g. by changing the evidence) will not influence the posterior probabilities in the other subsystem. This idea can be used to define more general Plefka expansions, where not just single node marginals but also marginal distributions of larger clusters are frozen. A similar theorem, stating that in such expansions no diagrams can appear where deleting any cluster of frozen nodes will cut the diagram in two, should apply.

11 Acknowledgements

We thank Jonathan Yedidia for explaining the concepts of the Gibbs free energy and answering many questions concerning the Plefka expansion. We are grateful to Geoffrey Hinton for writing part of the introduction of this paper and many interesting discussions on Boltzmann machines. David MacKay and Zoubin Ghahramani are also gratefully acknowledged for many insightful discussions on the topic of this paper. Finally, we thank the referees for their useful comments which improved the manuscript considerably.

References

[1] S. Amari. Information geometry of the EM and em algorithms for neural networks. *Neural Networks*, 8:1379–1408, 1995.

[2] B.J. Frey and D.J.C. MacKay. A revolution: Belief propagation in graphs with cycles. In *Advances in Neural Information Processing Systems*, volume 10, 1997.

[3] A. Georges and J.S. Yedidia. How to expand around mean-field theory using high-temperature expansions. *J. Phys A: Math. Gen.*, 24:2173–2192, 1991.

[4] G.E. Hinton and T.J. Sejnowski. Optimal perceptual inference. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, volume 352, 1983.

[5] G.E. Hinton and T.J. Sejnowski. Learning and relearning in Boltzmann machines. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations.* MIT Press, 1986.

[6] J.J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. In *Proceedings of the National Academy of Sciences*, volume 79, pages 2554–2558, 1982.

[7] M.I. Jordan, Z. Ghahramani, T. Jaakkola, and L.K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

[8] Y. Kabashima and D. Saad. Belief propagation vs. TAP for decoding corrupted messages. *Europhys. Lett.*, 44:668, 1998.

[9] H.J. Kappen and F.B. Rodriguez. Efficient learning in Boltzmann machines using linear response theory. *Neural Computation*, 10:1137–1156, 1998.

[10] M.A.R. Leisink and H.J. Kappen. Validity of TAP equations in neural networks. In *Proceedings of the International Conference on Artificial Neural Networks*, pages 425–430, 1999.

[11] D.J.C. MacKay, J.S. Yedidia, W.T. Freeman, and Y. Weiss. A conversation about the Bethe free energy and sum-product. http://www.inference.phy.cam.ac.uk/mackay/.

[12] R. McEliece, D. MacKay, and J. Cheng. Turbo decoding as an instance of Pearl's belief propagation algorithm. *IEEE J. Selected Areas in Communication, 1997*, 16:140–152, 1998.

[13] M. Mezard, G. Parisi, and M.A. Virasoro. *Spin Glass Theory and Beyond*. World Scientific Publishing Co., 1987.

[14] T. Morita. Variational principle for the distribution function of the effective field for the random Ising model in the Bethe approximaton. *Physica* **A**, 98:566–572, 1979.

[15] K. Murphy, Y. Weiss, and M. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 1999.

[16] H. Nishimori. *Statistical Physics of Spin Glasses and Information Processing*. Oxford, 2001.

[17] L. Onsager. Electric moments of molecules in liquids. *Journal of the American Chemical Society*, 58:1486–1493, 1936.

[18] M. Opper and D. Saad. *Advanced Mean Field Methods – Theory and Practice*. MIT Press, 2001.

[19] M. Opper and O. Winther. A mean field approach to bayes learning in large feed-forward neural networks. *Advances in Neural Information Processing Systems*, 9:225–331, 1997.

[20] M. Opper and O. Winther. From naive mean field theory to the TAP equations. In *Advanced Mean Field Methods – Theory and Practice*. MIT Press, 2001.

[21] G. Parisi. *Statistical Field Theory*. Perseus Books, 1988.

[22] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, San Mateo, California, 1988.

[23] C. Peterson and J. Anderson. A mean field theory learning algorithm for neural networks. *Complex Systems*, 1:995–1019, 1987.

[24] T. Plefka. Convergence condition of the TAP equations for infinite ranged Ising spin glass model. *Journal of Physics* **A**, 15:1971, 1982.

[25] Y.W. Teh and M. Welling. The unified propagation and scaling algorithm. In *Advances in Neural Information Processing Systems*, 2001.

[26] Y. Weiss. Correctness of local probability propagation in graphical models with loops. *Neural Computation*, 12:1–41, 2000.

[27] Y. Weiss. Comparing the mean field method and belief propagation for approximate inference in MRFs. In *Advanced Mean Field Methods – Theory and Practice*. MIT Press, 2001.

[28] Y. Weiss and W. Freeman. Correctness of belief propagation in gaussian graphical models of arbitrary topology. In *Advances in Neural Information Processing Systems*, volume 12, 1999.

[29] M. Welling and Y.W. Teh. Belief optimization for binary networks: a stable alternative to loopy belief propagation. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 554–561, 2001.

[30] W. Wiegerinck. Variational approximations between mean field theory and the junction tree algorithm. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 626–636, 2000.

[31] J.S. Yedidia. An idiosyncratic journey beyond mean field theory. In *Advanced Mean Field Methods – Theory and Practice*. MIT Press, 2001.

[32] J.S. Yedidia, W. Freeman, and Y. Weiss. Generalized belief propagation. In *Advances in Neural Information Processing Systems*, volume 13, 2000.

[33] A.L. Yuille. CCCP algorithms to minimize the Bethe and Kikuchi free energies: Convergent alternatives to belief propagation. *Neural Computation*, 14(7):1691–1722, 2002.