# On the uniqueness of loopy belief propagation fixed points

To appear in Neural Computation, 2004

Tom Heskes
SNN, University of Nijmegen
Geert Grooteplein 21, 6525 EZ, Nijmegen, The Netherlands
tom@snn.kun.nl

**Abstract**

We derive sufficient conditions for the uniqueness of loopy belief propagation fixed points. These conditions depend both on the structure of the graph and the strength of the potentials and naturally extend those for convexity of the Bethe free energy. We compare them with (a strengthened version of) conditions derived elsewhere for pairwise potentials. We discuss possible implications for convergent algorithms as well as for other approximate free energies.

## 1 Introduction

Loopy belief propagation is Pearl's belief propagation (Pearl 1988) applied to networks containing cycles. It can be used to compute approximate marginals in Bayesian networks and Markov random fields. Whereas belief propagation is exact only in special cases, e.g., for tree-structured (singly connected) networks with just Gaussian or just discrete nodes, loopy belief propagation empirically often leads to good performance (Murphy, Weiss & Jordan 1999, McEliece, MacKay & Cheng 1998). That is, the approximate marginals computed with loopy belief propagation are in many cases close to the exact marginals. In Gaussian graphical models, the means are guaranteed to coincide with the exact means (Weiss &Freeman 2001). The notion that fixed points of loopy belief propagation correspond to extrema of the so-called Bethe free energy (Yedidia, Freeman & Weiss 2001) is an important step in the theoretical understanding of this success and paved the road for interesting generalizations.

However, when applied to graphs with cycles, loopy belief propagation does not always converge. So-called double-loop algorithms have been proposed that do guarantee convergence (Yuille 2002, Teh & Welling 2002, Heskes, Albers & Kappen 2003), but are an order of magnitude slower than standard loopy belief propagation. It is generally believed that there is a close connection between (non)convergence of loopy belief propagation and (non)uniqueness of loopy belief propagation fixed points. More specifically, the working hypothesis is that uniqueness of a loopy belief propagation fixed point guarantees convergence of loopy belief propagation to this fixed point. The goal of this study then is to derive sufficient conditions for uniqueness. Such conditions are not only relevant from a theoretical point of view, but can also be used to derive faster algorithms and suggest different free energies, as will be discussed in Section 9.

# 2 Outline

Before getting into the mathematical details, we will first sketch the line of reasoning that will be followed in this article. It is inspired by the connection between fixed points of loopy belief propagation and extrema of the Bethe free energy: by studying the Bethe free energy we can learn about properties of loopy belief propagation.

The Bethe free energy is an approximation to the exact variational Gibbs-Helmholtz free energy. Both are concepts from (statistical) physics. Abstracting from the physical interpretation, the Gibbs-Helmholtz free energy is "just" a functional with a unique minimum, the argument of which corresponds to the exact probability distribution. However, the Gibbs-Helmholtz free energy is as intractable as the exact probability distribution. The idea is then to approximate the Gibbs-Helmholtz free energy in the hope that the minimum of such a tractable approximate free energy relates to the minimum of the exact free energy. Examples of such approximations are the mean-field free energy, the Bethe free energy, and the Kikuchi free energy. The connections between the Gibbs-Helmholtz free energy, Bethe free energy, and loopy belief propagation are reviewed in Section 3.

The Bethe free energy is a function of so-called pseudomarginals or beliefs. For the minimum of the Bethe free energy to make sense, these pseudomarginals have to be properly normalized as well as consistent. So, our starting point, the upper left corner in Figure 1, is a constrained minimization problem. In general, it is in fact a non-convex constrained minimization problem since the Bethe free energy is a non-convex function of the pseudomarginals (the constraints are linear in these pseudomarginals).

However, using the constraints on the pseudomarginals, it may be possible to rewrite the Bethe free energy in a form that *is* convex in the pseudomarginals.

When this is possible, we call the Bethe free energy "convex over the set of constraints" (Pakzad & Anantharam 2002). Now, if the Bethe free energy is convex over the set of constraints, we have, in combination with the linearity of the constraints, a convex constrained minimization problem. Convex constrained minimization problems have a unique solution (see e.g. (Luenberger 1984)), which explains link (d) in Figure 1.

Sufficient conditions for convexity over the set of constraints, link (b) in Figure 1, can be found in (Pakzad & Anantharam 2002) and (Heskes et al. 2003). They are (re)derived and discussed in Section 4. These conditions only depend on the structure of the graph, not on the (strength of the) potentials that make up the probability distribution defined over this graph. A corollary of these conditions, derived in Section 4.3, is that the Bethe free energy for a graph with a single loop is "just" convex over the set of constraints: with two or more connected loops the conditions fail (see also (McEliece & Yildirim 2003)).

Milder conditions for uniqueness, that do depend on the strength of the interactions, follow from the track on the righthand side of Figure 1. First, we note that non-convex constrained minimization of the Bethe free energy is equivalent to an unconstrained non-convex/concave minimax problem (Heskes 2002), link (a) in Figure 1. Convergent double-loop algorithms like CCCP (Yuille 2002) and faster variants thereof (Heskes et al. 2003) in fact solve such a minimax problem: the concave problem in the maximizing parameters (basically Lagrange multipliers) is solved by a message passing algorithm very similar to standard loopy belief propagation in the inner loop, where the outer loop changes the minimizing parameters (a remaining set of pseudomarginals) in the proper downward direction. The transformation from non-convex constrained minimization problem to an unconstrained non-convex/concave minimax problem is, in a particular setting relevant to this article, repeated in Section 5.1.

Rather than requiring the Bethe free energy to be convex (over the set of constraints), we then in Section 6 and Section 8 work towards conditions under which this minimax problem is convex/concave. These indeed depend on the strength of the potentials, defined in Section 7. These conditions can be considered the main result of this article. Link (c) follows from the observation, in Section 5.2, that the minimax problem corresponding to a Bethe free energy that is convex over the set of constraints has to be convex/concave.

As indicated by link (e), convex/concave minimax problems have a unique solution. This then also implies that the Bethe free energy has a unique extremum satisfying the constraints, which, since the Bethe free energy is bounded from below (see Section 5.3), has to be a minimum: link (f).

The concluding statement by link (g) in the lower right corner is, to the best of our knowledge, no more than a conjecture. We will discuss it in more detail in Section 9.

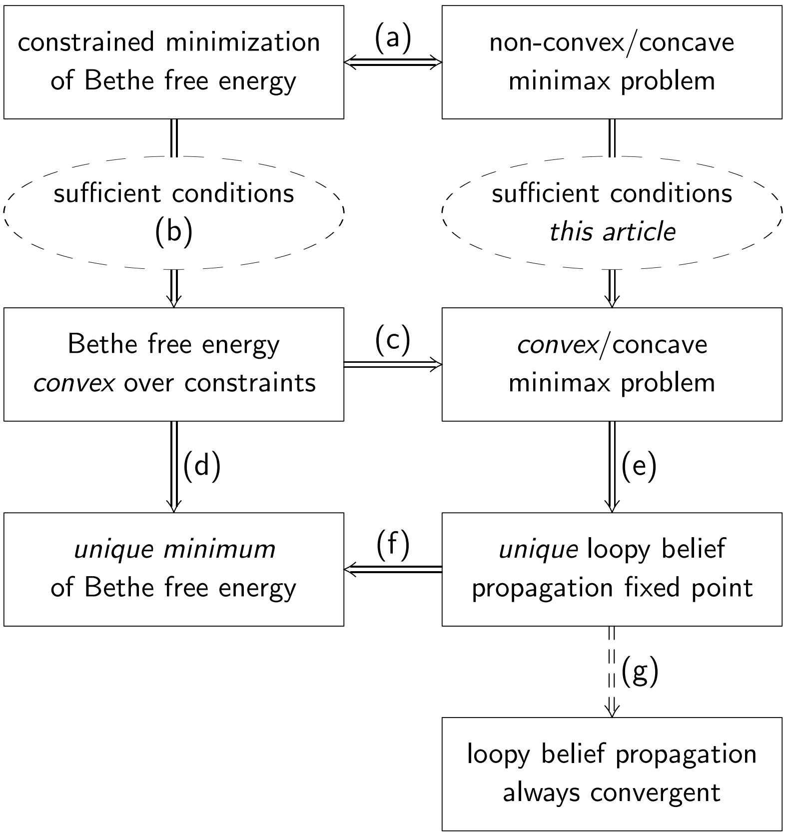



Figure 1: Layout of correspondences and implications. See text for details.

# 3 The Bethe free energy and loopy belief propagation

## 3.1 The Gibbs-Helmholtz free energy

The exact probability distribution in Bayesian networks and Markov random fields can be written in the factorized form

$$P_{\rm exact}(X) = \frac{1}{Z} \prod_\alpha \Psi_\alpha(X_\alpha) \; . \tag{1}$$

Here $\Psi_\alpha$ is a potential, some function of the potential subset $X_\alpha$, and $Z$ is an unknown normalization constant. Potential subsets typically overlap and they span the whole domain $X$. The convention that we adhere to in this article is that there are no potential subsets $X_\alpha$ and $X_{\alpha'}$ such that $X_{\alpha'}$ is fully subsumed by $X_\alpha$. The standard choice of a potential in a Bayesian network is a child with all its parents. We will further restrict ourselves to probabilistic models defined on discrete random variables, each of which runs over a finite number of states. The potentials are positive and finite.

The typical goal in Bayesian networks and Markov random fields is to compute the partition function $Z$ and/or marginals, for example

$$P_{\rm exact}(X_\alpha) = \sum_{X_{\backslash\alpha}} P_{\rm exact}(X) \; .$$

One way to do this is with the junction tree algorithm (Lauritzen & Spiegelhalter 1988). However, the junction tree algorithm scales exponentially with the size of the largest clique and may become intractable for complex models. The alternative is then to resort to approximate methods, that can be roughly divided into two categories: sampling approaches and deterministic approximations.

Most deterministic approximations derive from the so-called Gibbs-Helmholtz free energy

$$F(P) = -\sum_\alpha \sum_{X_\alpha} P(X_\alpha)\psi_\alpha(X_\alpha) + \sum_X P(X) \log P(X) \; ,$$

with shorthand $\psi \equiv \log \Psi$. Minimizing this variational free energy over the set $\mathcal{P}$ of all properly normalized probability distributions we get back the exact probability distribution (1) as the argument at the minimum and minus the log of the partition function as the value at the minimum:

$$P_{\rm exact} = \operatorname*{argmin}_{P \in \mathcal{P}} F(P) \quad \text{and} \quad -\log Z = \min_{P \in \mathcal{P}} F(P) \; .$$

Since the Gibbs-Helmholtz free energy is convex in $P$, the equality constraint (proper normalization) is linear, and the inequality constraints (non-negativity) are convex, this minimum is unique. By itself, we have not gained anything: the entropy may still be intractable to compute.

## 3.2 The Bethe free energy

The Bethe free energy is an approximation of the exact Gibbs-Helmholtz free energy. In particular, we approximate the entropy through

$$\sum_X P(X)\log P(X) \approx \sum_\alpha \sum_{X_\alpha} P(X_\alpha)\log P(X_\alpha) - \sum_\beta (n_\beta - 1)\sum_{x_\beta} P(x_\beta)\log P(x_\beta) ,$$

with $x_\beta$ a (super)node and $n_\beta = \sum_{\alpha\supset\beta} 1$: the number of potentials that contains node $x_\beta$. The second term follows from a discounting argument: without it we would overcount the entropy contributions on the overlap between the potential subsets. The (super)nodes $x_\beta$ are themselves subsets of the potential subsets, i.e.,

$$x_\beta \cap X_\alpha = \emptyset \text{ or } x_\beta \cap X_\alpha = x_\beta \quad \forall_{\alpha,\beta} ,$$

and partition the domain $X$, i.e.,

$$x_\beta \cap x_{\beta'} = \emptyset \quad \forall_{\beta,\beta'} \quad \text{and} \quad \bigcup_\beta x_\beta = X .$$

Typically the $x_\beta$ are taken to be single nodes and in the following we will therefore refer to them as such. For clarity of notation, we will indicate these nodes by $\beta$ and $x_\beta$ in lower case, to contrast them with the potentials $\alpha$ and potential subsets $X_\alpha$ in upper case.

Note that the Bethe free energy only depends on the marginals $P(X_\alpha)$ and $P(x_\beta)$. We replace minimization of the exact Gibbs-Helmholtz free energy over probability distributions by minimization of the Bethe free energy

$$\begin{aligned} F(Q_\alpha, Q_\beta) &= -\sum_\alpha \sum_{X_\alpha} Q_\alpha(X_\alpha)\psi_\alpha(X_\alpha) + \sum_\alpha \sum_{X_\alpha} Q_\alpha(X_\alpha)\log Q_\alpha(X_\alpha) \\ &\quad - \sum_\beta (n_\beta - 1)\sum_{x_\beta} Q_\beta(x_\beta)\log Q_\beta(x_\beta) , \end{aligned} \tag{2}$$

over sets of "pseudomarginals"[1] or beliefs $\{Q_\alpha, Q_\beta\}$. For this to make sense, these pseudomarginals have to be properly normalized as well as consistent, i.e.,[2]

$$\sum_{X_\alpha} Q_\alpha(X_\alpha) = 1 \quad \text{and} \quad Q_\alpha(x_\beta) = \sum_{X_{\alpha\setminus\beta}} Q_\alpha(X_\alpha) = Q_\beta(x_\beta) . \tag{3}$$

Let $\mathcal{Q}$ denote all subsets of consistent and properly normalized pseudomarginals, then our goal is to solve

$$\min_{\{Q_\alpha, Q_\beta\}\in\mathcal{Q}} F(Q_\alpha, Q_\beta) .$$

The hope is that the pseudomarginals at this minimum are accurate approximations to the exact marginals $P_{\text{exact}}(X_\alpha)$ and $P_{\text{exact}}(x_\beta)$.

[1] Terminology from e.g. (Wainwright, Jaakkola & Willsky 2002), used to indicate that there need not be a joint distribution that would yield such marginals.

[2] Strictly speaking we also have to take inequality constraints into account, namely those of the form $Q_\alpha(X_\alpha) \geq 0$. However, with the potentials being positive and finite the logarithmic terms in the free energy make sure that we never really have to worry about those, i.e., they never become "active". For convenience we will therefore not consider them any further.

### 3.3 Link with loopy belief propagation

For completeness and later reference we describe the link between the Bethe free energy and loopy belief propagation, as originally reported on in (Yedidia et al. 2001). It starts with the Lagrangian

$$\begin{aligned} L(Q_\alpha, Q_\beta, \lambda_{\alpha\beta}, \lambda_\alpha, \lambda_\beta) &= F(Q_\alpha, Q_\beta) + \sum_\beta \sum_{\alpha \supset \beta} \sum_{x_\beta} \lambda_{\alpha\beta}(x_\beta) \left[ Q_\beta(x_\beta) - Q_\alpha(x_\beta) \right] \\ &\quad + \sum_\alpha \lambda_\alpha \left[ 1 - \sum_{X_\alpha} Q_\alpha(X_\alpha) \right] + \sum_\beta \lambda_\beta \left[ 1 - \sum_\beta Q_\beta(x_\beta) \right] . \quad (4) \end{aligned}$$

At an extremum of the Bethe free energy satisfying the constraints all derivatives of $L$ are zero: the ones with respect to the Lagrange multipliers $\lambda$ give back the constraints; the ones with respect to the pseudomarginals $Q$ give an extremum of the Bethe free energy. Now, setting the derivatives with respect to $Q_\alpha$ and $Q_\beta$ to zero, we can solve for $Q_\alpha$ and $Q_\beta$ in terms of the Lagrange multipliers:

$$\begin{aligned} Q^*_\alpha(X_\alpha) &= \Psi_\alpha(X_\alpha) \exp \left[ \lambda_\alpha - 1 + \sum_{\beta \subset \alpha} \lambda_{\alpha\beta}(x_\beta) \right] \\ Q^*_\beta(x_\beta) &= \exp \left[ \frac{1}{n_\beta - 1} \left\{ 1 - \lambda_\beta + \sum_{\alpha \supset \beta} \lambda_{\alpha\beta}(x_\beta) \right\} \right] \end{aligned}$$

In terms of the"message" $\mu_{\beta \to \alpha}(x_\beta) \equiv \exp[\lambda_{\alpha\beta}(x_\beta)]$ from node $\beta$ to potential $\alpha$, the pseudomarginal $Q^*_\alpha(X_\alpha)$ reads

$$Q^*_\alpha(X_\alpha) \propto \Psi_\alpha(X_\alpha) \prod_{\beta \subset \alpha} \mu_{\beta \to \alpha}(x_\beta) \, , \quad (5)$$

where proper normalization yields the Lagrange multiplier $\lambda_\alpha$. With definition

$$\mu_{\alpha \to \beta}(x_\beta) \equiv \frac{Q^*_\beta(x_\beta)}{\mu_{\beta \to \alpha}(x_\beta)} \, , \quad (6)$$

the fixed-point equation for $Q^*_\beta(x_\beta)$ can, after some manipulations be written in the form

$$Q^*_\beta(x_\beta) \propto \prod_{\alpha \supset \beta} \mu_{\alpha \to \beta}(x_\beta) \, , \quad (7)$$

where again the Lagrange multiplier $\lambda_\beta$ follows from normalization. Finally, the constraint $Q^*_\alpha(x_\beta) = Q^*_\beta(x_\beta)$ in combination with (6) suggests the update

$$\mu_{\alpha \to \beta}(x_\beta) = \frac{Q^*_\alpha(x_\beta)}{\mu_{\beta \to \alpha}(x_\beta)} \, . \quad (8)$$

Equations (5) through (8) constitute the belief propagation equations. They can be summarized as follows. A pseudomarginal is the potential (just 1 for the nodes in the convention where no potentials are assigned to nodes) times its incoming messages; the outgoing message is the pseudomarginal divided by the incoming message. The scheduling of the messages is somewhat arbitrary. Loopy belief propagation can be "damped" by taking smaller steps. This damping is usually done in terms of the Lagrange multipliers, i.e., in the log-domain of the messages:

$$\log \mu^{\rm new}_{\alpha\to\beta}(x_\beta) = \log \mu_{\alpha\to\beta}(x_\beta) + \epsilon \left[ \{\log Q^*_\alpha(x_\beta) - \log \mu_{\beta\to\alpha}(x_\beta)\} - \log \mu_{\alpha\to\beta}(x_\beta) \right] . \tag{9}$$

Summarizing, loopy belief propagation is equivalent to fixed point iteration, where the fixed points are the zero derivatives of the Lagrangian.

# 4 Convexity of the Bethe free energy

## 4.1 Rewriting the Bethe free energy

Minimization of the Bethe free energy (2) under the constraints (3) is equivalent to solving a minimax problem on the Lagrangian (4), namely

$$\min_{Q_\alpha,Q_\beta} \max_{\lambda_{\alpha\beta},\lambda_\alpha,\lambda_\beta} L(Q_\alpha, Q_\beta, \lambda_{\alpha\beta}, \lambda_\alpha, \lambda_\beta) .$$

The ordering of the min and max operations is important here: to enforce the constraints, we first have to take the maximum. The min and max operations can be interchanged if we have a *convex* constrained minimization problem (Luenberger 1984). That is, the function to be minimized must be convex in its parameters, the equality constraints have to be linear, and the inequality constraints convex. In our case, the equality constraints are indeed linear and the inequality constraints enforcing non-negativity of the pseudomarginals indeed convex. However, the Bethe free energy (2) is clearly non-convex in its parameters $\{Q_\alpha, Q_\beta\}$. This is what makes it a difficult optimization problem.

Luckily the description (2) is not unique: any other form that can be constructed by substituting the constraints (3) is equally valid. Following (Pakzad & Anantharam 2002) we call the Bethe free energy "convex over the set of constraints" if, by making use of the constraints (3), we can rewrite it in a form that *is* convex in $\{Q_\alpha, Q_\beta\}$.

## 4.2 Conditions for convexity

The problem is with the term

$$S_\beta(Q_\beta) \equiv -\sum_{x_\beta} Q_\beta(x_\beta) \log Q_\beta(x_\beta) ,$$

which is concave in $Q_\beta$. Using the constraint $Q_\beta(x_\beta) = Q_\alpha(x_\beta)$, we can turn it into a functional that is convex in $Q_\alpha$ and $Q_\beta$ separately, but not necessarily jointly. That is, with the substitution $Q_\beta(x_\beta) = Q_\alpha(x_\beta)$ for any $\alpha \supset \beta$, the entropy and thus the Bethe free energy is convex in $Q_\alpha$ and in $Q_\beta$, but not necessarily in $\{Q_\alpha, Q_\beta\}$. However, if we add to $S_\beta(Q_\beta)$ a convex entropy contribution

$$-S_\alpha(Q_\alpha) \equiv \sum_{X_\alpha} Q_\alpha(X_\alpha) \log Q_\alpha(X_\alpha) \; ,$$

the combination of $-S_\alpha$ and $S_\beta$ is convex in $\{Q_\alpha, Q_\beta\}$, as the following Lemma, needed in the proof of Theorem 4.2 below, shows.

**Lemma 4.1**

$$\Delta_{\alpha\beta}(Q_\alpha, Q_\beta) \equiv \sum_{X_\alpha} Q_\alpha(X_\alpha) \log Q_\alpha(X_\alpha) - \sum_{x_\beta} Q_\alpha(x_\beta) \log Q_\beta(x_\beta)$$

*is convex in* $\{Q_\alpha, Q_\beta\}$.

**Proof** The matrix with second derivatives of $\Delta_{\alpha\beta}$ has components

$$\begin{aligned}
H(X_\alpha, X'_\alpha) &\equiv \frac{\partial^2 \Delta_{\alpha\beta}}{\partial Q_\alpha(X_\alpha) \partial Q_\alpha(X'_\alpha)} = \frac{1}{Q_\alpha(X_\alpha)} \delta_{X_\alpha, X'_\alpha} \\
H(X_\alpha, x'_\beta) &\equiv \frac{\partial^2 \Delta_{\alpha\beta}}{\partial Q_\alpha(X_\alpha) \partial Q_\beta(x'_\beta)} = -\frac{1}{Q_\beta(x_\beta)} \delta_{x_\beta, x'_\beta} \\
H(x_\beta, x'_\beta) &\equiv \frac{\partial^2 \Delta_{\alpha\beta}}{\partial Q_\beta(x_\beta) \partial Q_\beta(x'_\beta)} = -\frac{Q_\alpha(x_\beta)}{Q^2_\beta(x_\beta)} \delta_{x_\beta, x'_\beta} \; ,
\end{aligned}$$

where we note that $X_\alpha$ and $x_\beta$ should be interpreted as indices. Convexity requires that for any "vector" $\begin{pmatrix} R_\alpha(X_\alpha) & R_\beta(x_\beta) \end{pmatrix}$

$$\begin{aligned}
0 &\leq \begin{pmatrix} R_\alpha(X_\alpha) & R_\beta(x_\beta) \end{pmatrix} \begin{pmatrix} H(X_\alpha, X'_\alpha) & H(X_\alpha, x'_\beta) \\ H(x_\beta, X'_\alpha) & H(x'_\beta, x_\beta) \end{pmatrix} \begin{pmatrix} R_\alpha(X'_\alpha) \\ R_\beta(x'_\beta) \end{pmatrix} \\
&= \sum_{X_\alpha} \frac{R^2_\alpha(X_\alpha)}{Q_\alpha(X_\alpha)} - 2 \sum_{X_\alpha} \frac{R_\alpha(X_\alpha) R_\beta(x_\beta)}{Q_\beta(x_\beta)} + \sum_{x_\beta} \frac{Q_\alpha(x_\beta) R^2_\beta(x_\beta)}{Q^2_\beta(x_\beta)} \\
&= \sum_{X_\alpha} Q_\alpha(X_\alpha) \left[ \frac{R_\alpha(X_\alpha)}{Q_\alpha(X_\alpha)} - \frac{R_\beta(x_\beta)}{Q_\beta(x_\beta)} \right]^2 \; . \quad \blacksquare
\end{aligned}$$

The idea is then that the Bethe free energy is convex over the set of constraints if we have sufficient convex resources $Q_\alpha \log Q_\alpha$ to compensate for the concave $-Q_\beta \log Q_\beta$ terms. This can be formalized in the following theorem.

**Theorem 4.2** *The Bethe free energy is convex over the set of consistency constraints if there exists an "allocation matrix" $A_{\alpha\beta}$ between potentials $\alpha$ and nodes $\beta$ satisfying*

$$\begin{array}{lll} 1. & A_{\alpha\beta} \geq 0 \quad \forall_{\alpha,\beta\subset\alpha} & \textit{(positivity)} \\ 2. & \sum_{\beta\subset\alpha} A_{\alpha\beta} \leq 1 \quad \forall_{\alpha} & \textit{(sufficient amount of resources)} \\ 3. & \sum_{\alpha\supset\beta} A_{\alpha\beta} \geq n_\beta - 1 \quad \forall_{\beta} & \textit{(sufficient compensation)} \end{array} \tag{10}$$

**Proof** First of all, we note that we do not have to worry about the energy terms that are linear in $Q_\alpha$. In other words, to prove the theorem we can restrict ourselves to proving that minus the entropy

$$-S(Q) = -\left[\sum_\alpha S_\alpha(Q_\alpha) - \sum_\beta (n_\beta - 1) S_\beta(Q_\beta)\right]$$

is convex over the set of consistency constraints. The resulting operation is now a matter of resource allocation. For each concave contribution $(n_\beta - 1)S_\beta$ we have to find convex contributions $-S_\alpha$ to compensate for it. Let $A_{\alpha\beta}$ denote the "amount of resources" that we take from potential subset $\alpha$ to compensate for node $\beta$. Now, in shorthand notation and with a little bit of rewriting

$$\begin{aligned} -S(Q) &= -\left[\sum_\alpha S_\alpha - \sum_\beta (n_\beta - 1) S_\beta\right] \\ &= -\sum_\alpha \left(1 - \sum_{\beta\subset\alpha} A_{\alpha\beta} + \sum_{\beta\subset\alpha} A_{\alpha\beta}\right) S_\alpha \\ &\quad - \sum_\beta \left\{ - \sum_{\alpha\supset\beta} A_{\alpha\beta} + \sum_{\alpha\supset\beta} A_{\alpha\beta} - (n_\beta - 1) \right\} S_\beta \\ &= -\sum_\alpha \left(1 - \sum_{\beta\subset\alpha} A_{\alpha\beta}\right) S_\alpha - \sum_\alpha \sum_{\beta\subset\alpha} A_{\alpha\beta} [S_\alpha - S_\beta] \\ &\quad - \sum_\beta \left[\sum_{\alpha\supset\beta} A_{\alpha\beta} - (n_\beta - 1)\right] S_\beta \end{aligned}$$

Convexity of the first term is guaranteed if $1 - \sum_\beta A_{\alpha\beta} \geq 0$ (condition 2.), of the second term if $A_{\alpha\beta} \geq 0$ (condition 1. and Lemma 4.1), and of the third term if $\sum_\alpha A_{\alpha\beta} - (n_\beta - 1) \geq 0$ (condition 3.). ∎

This theorem is a special case of the one in (Heskes et al. 2003) for the more general Kikuchi free energy. Either one of the inequality signs in condition 2. and 3. of (10) can be replaced by an equality sign without any consequences.

## 4.3 Some implications

**Corollary 4.3** *The Bethe free energy for singly-connected graphs is convex over the set of constraints.*

**Proof** By construction. Choose one of the leaf nodes as the root $\beta^*$ and define

$$\begin{aligned} A_{\alpha\beta} &= 1 \qquad \text{iff } \beta \subset \alpha \text{ and } \beta \text{ closer to the root } \beta^* \text{ than any other } \beta' \subset \alpha \\ A_{\alpha\beta'} &= 0 \qquad \text{for all other } \beta'. \end{aligned}$$

Obviously, this choice of $A$ satisfies conditions 1. and 2. of (10). Arguing the other way around, for each $\beta \neq \beta^*$ there is just a single potential $\alpha \supset \beta$ that is closer to the root $\beta^*$ than $\beta$ itself (see the illustration in Figure 4.3) and thus there are precisely $n_\beta - 1$ contributions $A_{\alpha\beta} = 1$. The root itself gets $n_{\beta^*}$ contributions $A_{\alpha\beta^*} = 1$, which is even better. Hence also condition 3. is satisfied:

$$\sum_{\alpha \supset \beta} A_{\alpha\beta} = n_\beta - 1 \;\; \forall_{\beta \neq \beta^*} \quad \text{and} \quad \sum_{\alpha \supset \beta^*} A_{\alpha\beta^*} = n_{\beta^*} > n_{\beta^*} - 1 \,. \qquad \blacksquare$$

With the above construction of $A$ we are in a sense "eating up resources towards the root". At the root, we have one piece of resources left, which suggests that we can still enlarge the set of graphs for which convexity can be shown using Theorem 4.2. This leads to the next corollary.

**Corollary 4.4** *The Bethe free energy for graphs with a single loop is convex over the set of constraints.*

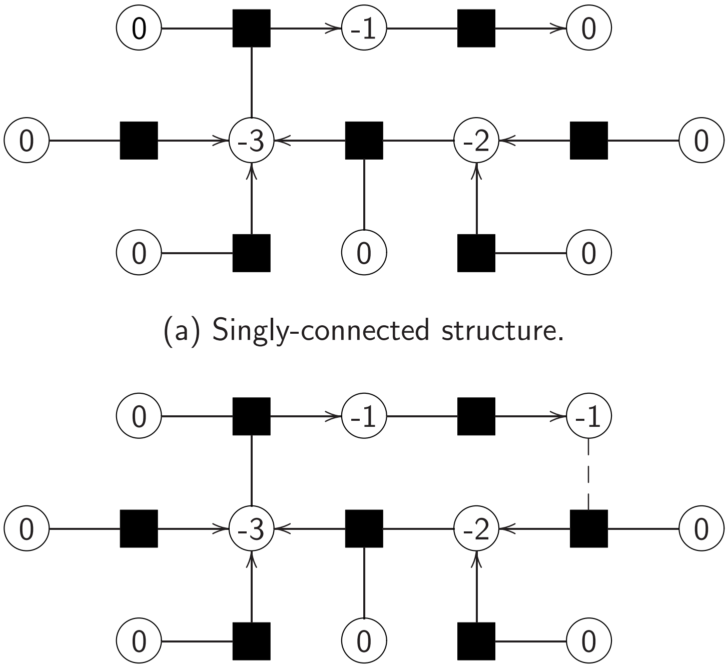
**Proof** Again by construction. Break the loop at one particular place, that is remove one node $\beta^*$ from a potential $\alpha^*$ such that a singly-connected structure is left. Construct a matrix $A$ as in the proof of Corollary 4.3, taking the node $\beta^*$ as the root. The matrix $A$ constructed in this way also just works for the graph with the closed loop since still

$$\sum_{\alpha \supset \beta} A_{\alpha\beta} = n_\beta - 1 \;\; \forall_{\beta \neq \beta^*} \quad \text{and now} \quad \sum_{\alpha \supset \beta^*} A_{\alpha\beta^*} = n_{\beta^*} - 1 \,. \qquad \blacksquare$$

It can be seen that this construction starts to fail as soon as we have two loops that are connected: with two connected loops we have insufficient positive resources to compensate for the negative entropy contributions.

## 4.4 Connection with other work

The same corollaries can be found in (Pakzad & Anantharam 2002, McEliece & Yildirim 2003). Furthermore, the conditions in Theorem 4.2 are very similar to the ones stated in (Pakzad & Anantharam 2002), which for the Bethe free energy boil down to the following.




(a) Singly-connected structure.

(b) Single-loop structure.

Figure 2: Illustration of the construction of an allocation matrix satisfying all convexity constraints for singly-connected (a) and single-loop structures (b). Neglecting the arrows and dashes, each graph correspond to a factor graph (Kschischang et al. 2001), where dark boxes refer to potentials and circles to nodes. The numbers within the circles give the corresponding "overcounting numbers", for the Bethe free energy $1 - n_\beta$ with $n_\beta$ the number of neighboring potentials. The arrows, pointing from potentials $\alpha$ to nodes $\beta$, visualize the allocation matrix $A$ with $A_{\alpha\beta} = 1$ if there is an arrow and $A_{\alpha\beta} = 0$ otherwise. As can be seen, for each potential there is precisely one outgoing arrow, pointing at the node closest to the root, chosen to be the node in the upper right corner of the graph. In the singly-connected structure (a), all non-root nodes have precisely $n_\beta - 1$ incoming arrows, just sufficient to compensate the overcounting number $1 - n_\beta$. The root node itself has one incoming arrow, which it does not really need. In the structure with the single loop (b), we open the loop by breaking the dashed link and construct the allocation matrix for the corresponding singly-connected structure. This allocation matrix works for the single-loop structure as well, because now the incoming arrow at the "root" is just sufficient to compensate for the negative overcounting number resulting from the extra link closing the loop.

**Theorem 4.5** *Adapted from Theorem 1 in (Pakzad & Anantharam 2002). The Bethe free energy is convex for the set of constraints if for any set of nodes B we have*

$$\sum_{\beta \in B} (1 - n_\beta) + \sum_{\alpha \in \pi(B)} 1 \geq 0 \,, \tag{11}$$

*where $\pi(B) \equiv \{\alpha : \exists \beta \in B; \beta \subset \alpha\}$ denotes the "parent" set of $B$, i.e., those potential subsets that include at least one node in $B$.*

**Proposition 4.6** *The conditions in Theorem 4.2 and those in Theorem 4.5 are equivalent.*

**Proof** Let us first suppose that there does exist an allocation matrix $A_{\alpha\beta}$ satisfying the conditions (10). Then for any set $B$,

$$\sum_{\beta \in B} (n_\beta - 1) \leq \sum_{\beta \in B} \sum_{\alpha \supset \beta} A_{\alpha\beta} \leq \sum_{\alpha \in \pi(B)} \sum_{\beta \subset \alpha} A_{\alpha\beta} \leq \sum_{\alpha \in \pi(B)} 1 \,,$$

where the inequalities follow from conditions 3., 1., and 2. in (10), respectively. In other words, validity of the conditions in Theorem 4.2 implies the validity of those in Theorem 4.5.

Next let us suppose that the conditions in Theorem 4.2 fail. Above we have seen that this can happen if and only if the graph contains at least one connected component with two connected loops. But then condition (11) is violated as well when we take for $B$ the set of all nodes within such a component.

Since validity implies validity and violation implies violation, the conditions must be equivalent. ∎

Graphical models with a single loop have been studied in detail in (Weiss 2000), yielding important theoretical results (e.g., correctness of maximum *a posteriori* assignments). These results are derived by "unwrapping" the single loop into an infinite tree. This argument also breaks down as soon as soon is there is more than a single loop. It might be interesting to find out whether there is a deeper connection between this unwrapping argument and the convexity of the Bethe free energy.

In summary, we have given conditions for the Bethe free energy to have a unique extremum satisfying the constraints. From the connection between the extrema of the Bethe free energy and fixed points of loopy belief propagation, it then follows that loopy belief propagation has a unique fixed point when these conditions are satisfied. These conditions fail as soon as the structure of the graph contains two connected loops.

The conditions for convexity of the Bethe free energy only depend on the structure of the graph: the potentials $\Psi_\alpha(X_\alpha)$ do not play any role. These potentials only appear in the energy term that is linear in the pseudomarginals and thus does not affect the convexity argument. Consequently, adding a "fake link" with potential $\Psi_\alpha(X_\alpha) = 1$ can change the validity of the conditions, whereas it has no effect on the loopy belief propagation updates. Even if we would manage to find more interesting (i.e., milder) conditions for convexity over the set of constraints[3], the above impact of fake links would never disappear. In the following we will therefore dig a little deeper to arrive at milder conditions that do take into account (the strength of) the potentials.

# 5 The dual formulation

## 5.1 From Lagrangian to dual

As we have seen, fixed points of loopy belief propagation are in one-to-one correspondence with zero derivatives of the Lagrangian. So, if we manage to find conditions under which these zero derivatives have a unique solution, then for the same conditions loopy belief propagation has a unique fixed point. In the following, we will work with a Lagrangian slightly different from (4). First, we substitute the constraint $Q_\alpha(x_\beta) = Q_\beta(x_\beta)$ to write the Bethe free energy in the "more convex" form

$$\begin{aligned} F(Q_\alpha, Q_\beta) \;\; &= \;\; -\sum_\alpha \sum_{X_\alpha} Q_\alpha(X_\alpha)\psi_\alpha(X_\alpha) + \sum_\alpha \sum_{X_\alpha} Q_\alpha(X_\alpha) \log Q_\alpha(X_\alpha) \\ &\quad - \sum_\beta \sum_{\alpha \supset \beta} A_{\alpha\beta} \sum_{x_\beta} Q_\alpha(x_\beta) \log Q_\beta(x_\beta) \,, \end{aligned} \tag{12}$$

where the allocation matrix $A_{\alpha\beta}$ can be any matrix that satisfies

$$\sum_{\alpha \supset \beta} A_{\alpha\beta} = n_\beta - 1 \,. \tag{13}$$

And secondly, we express the consistency constraints from (3) in terms of the potential pseudomarginals $Q_\alpha$ alone. This then yields

$$\begin{aligned} L(Q_\alpha, Q_\beta, \lambda_{\alpha\beta}, \lambda_\alpha) \;\; &= \;\; -\sum_\alpha \sum_{X_\alpha} Q_\alpha(X_\alpha)\psi_\alpha(X_\alpha) + \sum_\alpha \sum_{X_\alpha} Q_\alpha(X_\alpha) \log Q_\alpha(X_\alpha) \\ &\quad - \sum_\alpha \sum_{\beta \subset \alpha} A_{\alpha\beta} \sum_{x_\beta} Q_\alpha(x_\beta) \log Q_\beta(x_\beta) \\ &\quad + \sum_\beta \sum_{\alpha \supset \beta} \sum_{x_\beta} \lambda_{\alpha\beta}(x_\beta) \left[ \frac{1}{n_\beta - 1} \sum_{\alpha' \supset \beta} A_{\alpha'\beta} Q_{\alpha'}(x_\beta) - Q_\alpha(x_\beta) \right] \\ &\quad + \sum_\alpha \lambda_\alpha \left[ 1 - \sum_{X_\alpha} Q_\alpha(X_\alpha) \right] + \sum_\beta (n_\beta - 1) \left[ \sum_{x_\beta} Q_\beta(x_\beta) - 1 \right] \,, \end{aligned} \tag{14}$$

[3] We would like to conjecture that this is not possible, i.e., that the conditions in Theorem 4.2 are not only sufficient but also necessary to prove convexity of the Bethe free energy over the set of consistency constraints. Note that this would not imply that we need these conditions to guarantee the uniqueness of fixed points, since for that convexity by itself is sufficient, not necessary.

Note that the constraint $Q_\beta(x_\beta) = Q_\alpha(x_\beta)$ as well as its normalization is no longer incorporated with Lagrange multipliers, but follows when we take the minimum with respect to $Q_\beta$. It is easy to check that the fixed-point equations of loopy belief propagation still follow by setting the derivatives of the Lagrangian (14) to zero.

Although the Bethe free energy and thus the Lagrangian (14) may not be convex in $\{Q_\alpha, Q_\beta\}$, they are convex in $Q_\alpha$ and $Q_\beta$ separately. Therefore we can interchange the minimum over the pseudomarginals $Q_\alpha$ and the maximum over the Lagrange multipliers, as long as we leave the minimum over $Q_\beta$ as the final operation:[4]

$$\min_{Q_\alpha, Q_\beta} \max_{\lambda_{\alpha\beta}, \lambda_\alpha} L(Q_\alpha, Q_\beta, \lambda_{\alpha\beta}, \lambda_\alpha) = \min_{Q_\beta} \max_{\lambda_{\alpha\beta}, \lambda_\alpha} \min_{Q_\alpha} L(Q_\alpha, Q_\beta, \lambda_{\alpha\beta}, \lambda_\alpha) \, .$$

Rewriting

$$\sum_\beta \sum_{\alpha \supset \beta} \sum_{x_\beta} \lambda_{\alpha\beta}(x_\beta) \left[ \frac{1}{n_\beta - 1} \sum_{\alpha' \supset \beta} A_{\alpha'\beta} Q_{\alpha'}(x_\beta) - Q_\alpha(x_\beta) \right] = \\ - \sum_\alpha \sum_{\beta \subset \alpha} \sum_{x_\beta} \bar{\lambda}_{\alpha\beta}(x_\beta) Q_\alpha(x_\beta) \, ,$$

with

$$\bar{\lambda}_{\alpha\beta}(x_\beta) \equiv \lambda_{\alpha\beta}(x_\beta) - \frac{1}{n_\beta - 1} \sum_{\alpha' \supset \beta} A_{\alpha'\beta} \lambda_{\alpha'\beta}(x_\beta) \, ,$$

we can easily solve for the minimum with respect to $Q_\alpha$:

$$Q^*_\alpha(X_\alpha) = \Psi_\alpha(X_\alpha) \exp \left[ \lambda_\alpha - 1 + \sum_{\beta \subset \alpha} \left\{ A_{\alpha\beta} \log Q_\beta(x_\beta) + \bar{\lambda}_{\alpha\beta}(x_\beta) \right\} \right] \, . \tag{15}$$

Plugging this into the Lagrangian we obtain the "dual"

$$\begin{aligned} G(Q_\beta, \lambda_{\alpha\beta}, \lambda_\alpha) \;&\equiv\; L(Q^*_\alpha, Q_\beta, \lambda_{\alpha\beta}, \lambda_\alpha) \\ &=\; - \sum_\alpha \sum_{X_\alpha} \Psi_\alpha(X_\alpha) \exp \left[ \lambda_\alpha - 1 + \sum_{\beta \subset \alpha} \left\{ A_{\alpha\beta} \log Q_\beta(x_\beta) + \bar{\lambda}_{\alpha\beta}(x_\beta) \right\} \right] \\ &+\; \sum_\alpha \lambda_\alpha + \sum_\beta (n_\beta - 1) \left[ \sum_{x_\beta} Q_\beta(x_\beta) - 1 \right] \, . \end{aligned} \tag{16}$$

Next, we find for the maximum with respect to $\lambda_\alpha$

$$\exp\left[1 - \lambda^*_\alpha\right] = \sum_{X_\alpha} \Psi_\alpha(X_\alpha) \exp \left[ \sum_{\beta \subset \alpha} \left\{ A_{\alpha\beta} \log Q_\beta(x_\beta) + \bar{\lambda}_{\alpha\beta}(x_\beta) \right\} \right] \equiv Z^*_\alpha \, , \tag{17}$$

[4]In principle we could also first take the minimum over $Q_\beta$ and leave the minimum over $Q_\alpha$, but this does not seem to lead to any useful results.

where we have to keep in mind that $Z_\alpha^*$ by itself, like $Q_\alpha^*$ is a function of the remaining pseudomarginals $Q_\beta$ and Lagrange multipliers $\lambda_{\alpha\beta}$. Substituting this solution into the dual we arrive at

$$G(Q_\beta, \lambda_{\alpha\beta}) \equiv G(Q_\beta, \lambda_{\alpha\beta}, \lambda_\alpha^*) = -\sum_\alpha \log Z_\alpha^* + \sum_\beta (n_\beta - 1) \left[ \sum_{x_\beta} Q_\beta(x_\beta) - 1 \right] . \quad (18)$$

Let us pause here for a moment and reflect on what we have done sofar. The Lagrangian (14), being convex in $Q_\alpha$, has a unique minimum in $Q_\alpha$ (given all other parameters fixed), which is also the only extremum. It happens to be relatively straightforward to express the value at this minimum in terms of the remaining parameters and then also to find the optimal (maximal) $\lambda_\alpha^*$. Plugging these values into the Lagrangian (14), we have not lost anything. That is, zero derivatives of the Lagrangian (14) are still in one-to-one correspondence with zero derivatives of the dual (18) and thus with fixed points of loopy belief propagation.

## 5.2 Recovering the convexity conditions (1)

To find a minimum of the Bethe free energy satisfying the constraints (3), we first have to take the maximum of the dual (18) over the remaining Lagrange multipliers $\lambda_{\alpha\beta}$ and then the minimum over the remaining pseudomarginals $Q_\beta$. The duality theorem, a standard result from constrained optimization (see e.g. (Luenberger 1984)) tells us that the dual $G$ is concave in the Lagrange multipliers. The remaining question is then whether the dual is convex in $Q_\beta$. If it is, we have a convex/concave minimax problem, which is guaranteed to have a unique solution.

The link (c) in Figure 1 follows from the following proposition.

**Proposition 5.1** *Convexity of the Bethe free energy (12) in $\{Q_\alpha, Q_\beta\}$ implies convexity of the dual (18) in $Q_\beta$.*

**Proof** First we note that the minimum of a convex function over some of its parameters is convex in its remaining parameters. In obvious one-dimensional notation, with $y^*(x) \equiv \underset{y}{\operatorname{argmin}} f(x,y)$,

$$\begin{aligned} f(x+\delta, y^*(x+\delta)) + f(x-\delta, y^*(x-\delta)) &\geq 2f(x, (y^*(x+\delta) + y^*(x-\delta))/2) \\ &\geq 2f(x, y^*(x)) \, , \end{aligned}$$

where the first inequality follows from the convexity of $f$ in $\{x, y\}$ and the second inequality from $y^*(x)$ being the unique minimum of $f(x,y)$. Therefore, the dual (16) is convex in $Q_\beta$ when the Lagrangian (14) and thus the Bethe free energy (12) is convex in $\{Q_\alpha, Q_\beta\}$. Furthermore, from the duality theorem the dual (16) is concave in the Lagrange multipliers $\{\lambda_{\alpha\beta}, \lambda_\alpha\}$. Next we note that the maximum of a convex/concave function over its maximizing parameters is again convex: with $y^*(x) \equiv \underset{y}{\operatorname{argmax}} f(x,y)$,

$$
\begin{aligned}
f(x+\delta, y^*(x+\delta)) + f(x-\delta, y^*(x-\delta)) &\geq f(x+\delta, y^*(x)) + f(x-\delta, y^*(x)) \\
&\geq 2f(x, y^*(x)),
\end{aligned}
$$

where the first inequality follows from $y^*(x \pm \delta)$ being the unique maximum of $f(x \pm \delta, y)$ and the second inequality from the convexity of $f(x,y)$ in $x$. Hence the dual (18) must still be convex in $Q_\beta$. ∎

For now we did not gain nor lose anything in comparison with the conditions for Theorem 4.2. However, the inequalities in the above proof suggest a little space that will lead to milder conditions for the uniqueness of fixed points.

## 5.3 Boundedness of the Bethe free energy

For completeness and to support the link (f) in Figure 1 we will here first prove that the Bethe free energy is bounded from below. The following theorem can be considered a special case of the one stated in (Minka 2001) on the Bethe free energy for expectation propagation, a generalization of (loopy) belief propagation.

**Theorem 5.2** *If all potentials are bounded from above, i.e., $\Psi_\alpha(X_\alpha) \leq \Psi_{\max}$ for all $\alpha$ and $X_\alpha$, the Bethe free energy is bounded from below on the set of constraints.*

**Proof** It is sufficient to prove that the function $G(Q_\beta) \equiv \max_{\lambda_{\alpha\beta}} G(Q_\beta, \lambda)$ is bounded from below for a particular choice of $A_{\alpha\beta}$ satisfying (13). Considering $A_{\alpha\beta} = \frac{n_\beta - 1}{n_\beta}$, we then have

$$
\begin{aligned}
G(Q_\beta) &\geq -\sum_\alpha \log \sum_{X_\alpha} \Psi_\alpha(X_\alpha) \exp\left[\sum_{\beta \subset \alpha} \frac{n_\beta - 1}{n_\beta} \log Q_\beta(x_\beta)\right] \\
&\quad + \sum_\beta (n_\beta - 1)\left[\sum_{x_\beta} Q_\beta(x_\beta) - 1\right] \\
&\geq -\sum_\alpha \sum_{\beta \subset \alpha} \frac{n_\beta - 1}{n_\beta} \log \sum_{X_\alpha} \Psi_\alpha(X_\alpha) Q_\beta(x_\beta) \\
&\quad + \sum_\beta (n_\beta - 1)\left[\sum_{x_\beta} Q_\beta(x_\beta) - 1\right] \\
&\geq -\sum_\alpha \sum_{\beta \subset \alpha} \frac{n_\beta - 1}{n_\beta} \log\left[\sum_{X_{\alpha \setminus \beta}} \Psi_{\max}\right]
\end{aligned}
$$

$$
\begin{aligned}
& + \sum_{\beta} (n_\beta - 1) \left[ -\log \sum_{x_\beta} Q_\beta(x_\beta) + \sum_{x_\beta} Q_\beta(x_\beta) - 1 \right] \\
\geq \quad & -\sum_{\alpha} \sum_{\beta \subset \alpha} \frac{n_\beta - 1}{n_\beta} \log \left[ \sum_{X_{\alpha \backslash \beta}} \Psi_{\max} \right] ,
\end{aligned}
$$

where the first inequality follows by substituting the choice $\lambda_{\alpha\beta}(x_\beta) = 0$ for all $\alpha$, $\beta$, and $x_\beta$ in $G(Q_\beta, \lambda_{\alpha\beta})$, the second from the concavity of the function $y^{\frac{n_\beta - 1}{n_\beta}}$, and the third from the upper bound on the potentials. ∎

# 6 Towards better conditions

## 6.1 The Hessian

The next step is to compute the Hessian, i.e., the second derivative of the dual with respect to the pseudomarginals $Q_\beta$. The first derivative yields

$$
\frac{\partial G}{\partial Q_\beta(x_\beta)} = -\sum_{\alpha \supset \beta} A_{\alpha\beta} \frac{Q^*_\alpha(x_\beta)}{Q_\beta(x_\beta)} + (n_\beta - 1) ,
$$

which is immediate from the Lagrangian (14). To compute the matrix of second derivatives

$$
H_{\beta\beta'}(x_\beta, x'_{\beta'}) \equiv \frac{\partial^2 G}{\partial Q_\beta(x_\beta) \partial Q_{\beta'}(x'_{\beta'})}
$$

we make use of

$$
\frac{\partial Q^*_\alpha(x_\beta)}{\partial Q_{\beta'}(x'_{\beta'})} = A_{\alpha\beta'} \frac{Q^*_\alpha(x_\beta, x'_{\beta'}) - Q^*_\alpha(x_\beta) Q^*_\alpha(x'_{\beta'})}{Q_{\beta'}(x'_{\beta'})} ,
$$

where both $\beta$ and $\beta'$ should be a subset of $\alpha$ and with convention $Q^*_\alpha(x_\beta, x_\beta) = Q^*_\alpha(x_\beta)$ and $Q^*_\alpha(x_\beta, x'_\beta) = 0$ if $x_\beta \neq x'_\beta$. Here the first term follows from the differentation of (15) and the second term from the normalization as in (17). Distinguishing between $\beta = \beta'$ and $\beta \neq \beta'$, we then have

$$
\begin{aligned}
H_{\beta\beta}(x_\beta, x'_\beta) &= \sum_{\alpha \supset \beta} A_{\alpha\beta}(1 - A_{\alpha\beta}) \frac{Q^*_\alpha(x_\beta)}{Q^2_\beta(x_\beta)} \delta_{x_\beta, x'_\beta} + \sum_{\alpha \supset \beta} A^2_{\alpha\beta} \frac{Q^*_\alpha(x_\beta) Q^*_\alpha(x'_\beta)}{Q_\beta(x_\beta) Q_\beta(x'_\beta)} \\
H_{\beta\beta'}(x_\beta, x'_{\beta'}) &= -\sum_{\alpha \supset \{\beta, \beta'\}} A_{\alpha\beta} A_{\alpha\beta'} \frac{Q^*_\alpha(x_\beta, x'_{\beta'}) - Q^*_\alpha(x_\beta) Q^*_\alpha(x'_{\beta'})}{Q_\beta(x_\beta) Q_{\beta'}(x'_{\beta'})} \quad \text{for} \quad \beta' \neq \beta ,
\end{aligned}
$$

where $\delta_{x_\beta, x'_\beta} = 1$ if and only if $x_\beta = x'_\beta$. Here it should be noted that both $\beta$ and $x_\beta$ play the role of indices, i.e., $x_\beta$ should not be mistaken for a variable or parameter. The parameters are still the (tables with) Lagrange multipliers $\lambda_{\alpha\beta}$ and pseudomarginals $Q_\beta$.

The goal is now to find conditions under which this Hessian is positive (semi) definite for any setting of the parameters $\{Q_\beta, \lambda_{\alpha\beta}\}$. That is, conditions that guarantee

$$K \equiv \sum_{\beta,\beta'} \sum_{x_\beta, x_{\beta'}} S_\beta(x_\beta) H_{\beta\beta'}(x_\beta, x_{\beta'}) S_{\beta'}(x_{\beta'}) \geq 0 \,,$$

for any choice of the "vector" $S$ with elements $S_\beta(x_\beta)$. Straightforward manipulations yield

$$\begin{aligned}
&\sum_{\beta,\beta'} \sum_{x_\beta, x_{\beta'}} S_\beta(x_\beta) H_{\beta\beta'}(x_\beta, x_{\beta'}) S_{\beta'}(x_{\beta'}) = && (K) \\
&= \sum_\alpha \sum_{\beta \subset \alpha} \sum_{x_\beta} A_{\alpha\beta}(1 - A_{\alpha\beta}) Q^*_\alpha(x_\beta) R^2_\beta(x_\beta) && (K_1) \\
&+ \sum_\alpha \sum_{\{\beta,\beta'\} \subset \alpha} \sum_{x_\beta, x'_{\beta'}} A_{\alpha\beta} A_{\alpha\beta'} Q^*_\alpha(x_\beta) Q^*_\alpha(x'_{\beta'}) R_\beta(x_\beta) R_{\beta'}(x'_{\beta'}) && (K_2) \\
&- \sum_\alpha \sum_{\substack{\{\beta,\beta'\} \subset \alpha \\ \beta' \neq \beta}} \sum_{x_\beta, x'_{\beta'}} A_{\alpha\beta} A_{\alpha\beta'} Q^*_\alpha(x_\beta, x'_{\beta'}) R_\beta(x_\beta) R_{\beta'}(x'_{\beta'}) && (K_3)
\end{aligned}$$

where $R_\beta(x_\beta) \equiv S_\beta(x_\beta)/Q_\beta(x_\beta)$.

## 6.2 Recovering the convexity conditions (2)

Let us first see how we get back the conditions for convexity of the Bethe free energy (12). Since

$$K_2 = \sum_\alpha \left[ \sum_{\beta \subset \alpha} \sum_{x_\beta} A_{\alpha\beta} Q^*_\alpha(x_\beta) R_\beta(x_\beta) \right]^2 \geq 0$$

and[5]

$$\begin{aligned}
K_3 &= \sum_\alpha \sum_{\substack{\{\beta,\beta'\} \subset \alpha \\ \beta' \neq \beta}} \sum_{x_\beta, x'_{\beta'}} A_{\alpha\beta} A_{\alpha\beta'} Q^*_\alpha(x_\beta, x'_{\beta'}) \times \\
&\qquad \left\{ \frac{1}{2} \left[ R_\beta(x_\beta) - R_{\beta'}(x'_{\beta'}) \right]^2 - \frac{1}{2} R^2_\beta(x_\beta) - \frac{1}{2} R^2_{\beta'}(x'_{\beta'}) \right\} \\
&\geq \sum_\alpha \sum_{\beta \subset \alpha} \sum_{x_\beta} A_{\alpha\beta} \left( \sum_{\beta' \subset \alpha} A_{\alpha\beta'} - A_{\alpha\beta} \right) Q^*_\alpha(x_\beta) R^2_\beta(x_\beta) \,,
\end{aligned} \tag{19}$$

we have

$$K = K_1 + K_2 + K_3 \geq \sum_\alpha \sum_{\beta \subset \alpha} \sum_{x_\beta} A_{\alpha\beta} \left( 1 - \sum_{\beta' \subset \alpha} A_{\alpha\beta'} \right) Q^*_\alpha(x_\beta) R^2_\beta(x_\beta) \,.$$

[5]This step is in fact equivalent to Gerschgorin theorem for bounding the eigenvalues of a matrix.

That is, sufficient conditions for $K$ to be non-negative are

$$A_{\alpha\beta} \geq 0 \quad \forall_{\alpha,\beta\subset\alpha} \quad \text{and} \quad \sum_{\beta\subset\alpha} A_{\alpha\beta} \leq 1 \quad \forall_\alpha \; ,$$

precisely the conditions for Theorem 4.2.

## 6.3 Fake interactions

While discussing the conditions for convexity of the Bethe free energy, we noticed that adding a “fake interaction”, e.g., a constant potential, can change the validity of the conditions. We will see that here this is *not* the case and these fake interactions drop out as we would expect them to.

Suppose that we have a fake interaction $\Psi_\alpha(X_\alpha) = 1$. From the solution (15) it follows that the pseudomarginal $Q^*_\alpha(X_\alpha)$ factorizes[6]:

$$Q^*_\alpha(x_\beta, x'_{\beta'}) = Q^*_\alpha(x_\beta) Q^*_\alpha(x'_{\beta'}) \qquad \forall_{\{\beta,\beta'\}\subset\alpha} \; .$$

Consequently, the terms involving $\alpha$ in $K_3$ cancel with those in $K_2$, which is most easily seen when we combine $K_2$ and $K_3$ in a different way:

$$K_2 + K_3 = \sum_\alpha \sum_{\beta\subset\alpha} \sum_{x_\beta, x'_{\beta'}} A^2_{\alpha\beta} Q^*_\alpha(x_\beta) Q^*_\alpha(x'_\beta) R_\beta(x_\beta) R_\beta(x'_\beta) \tag{$\tilde{K}_2$}$$

$$- \sum_\alpha \sum_{\substack{\{\beta,\beta'\}\subset\alpha \\ \beta'\neq\beta}} \sum_{x_\beta, x'_{\beta'}} A_{\alpha\beta} A_{\alpha\beta'} \left[ Q^*_\alpha(x_\beta, x'_{\beta'}) - Q^*_\alpha(x_\beta) Q^*_\alpha(x'_{\beta'}) \right] R_\beta(x_\beta) R_{\beta'}(x'_{\beta'}) \; . \tag{$\tilde{K}_3$}$$

This leaves us with the weaker requirement (from $K_1$) $A_{\alpha\beta}(1 - A_{\alpha\beta}) \geq 0$ for all $\beta \subset \alpha$. The best choice is then to take $A_{\alpha\beta} = 1$, which turns condition 3. of (10) into

$$\sum_{\substack{\alpha'\supset\beta \\ \alpha'\neq\alpha}} A_{\alpha'\beta} + 1 \geq n_\beta - 1 \; .$$

The net effect is then equivalent to ignoring the interaction, reducing the number of neighboring potentials $n_\beta$ by 1 for all $\beta$ that are part of the fake interaction $\alpha$.

We have seen how we get milder and thus better conditions when there is effectively no interaction. Motivated by this “success” we will work towards conditions that take into account the *strength* of the interactions. Our starting point will be the above decomposition in $\tilde{K}_2$ and $\tilde{K}_3$ where, since $\tilde{K}_2 \geq 0$ anyways, we will concentrate on $\tilde{K}_3$.

[6] The exact marginal $P_{\text{exact}}(X_\alpha)$ need not factorize: this is really a consequence of the locality assumptions behind loopy belief propagation and the Bethe free energy.

# 7 The strength of a potential

## 7.1 Bounding the correlations

The crucial observation, that will allow us to obtain milder and thus better conditions for the uniqueness of a fixed point, is the following lemma. It bounds the term between brackets in $\tilde{K}_3$ such that we can again combine this bound with the (positive) term $K_1$. However, before we will get to that, we will take some time to introduce and derive properties of the "strength" of a potential.

**Lemma 7.1** *Two-node correlations of loopy belief marginals obey the bound*

$$Q^*_\alpha(x_\beta, x'_{\beta'}) - Q^*_\alpha(x_\beta)Q^*_\alpha(x'_{\beta'}) \leq \sigma_\alpha Q^*_\alpha(x_\beta, x'_{\beta'}) \quad \forall_{\substack{\{\beta,\beta'\}\subset\alpha \\ \beta'\neq\beta}} \forall_{x_\beta, x'_{\beta'}} , \tag{20}$$

*with the "strength"* $\sigma_\alpha$ *a function of the potential* $\psi_\alpha(X_\alpha) \equiv \log \Psi_\alpha(X_\alpha)$ *only:*

$$\begin{aligned} \sigma_\alpha &= 1 - \exp(-\omega_\alpha) \quad \textit{with} \\ \omega_\alpha &\equiv \max_{X_\alpha, \hat{X}_\alpha} \left[ \psi_\alpha(X_\alpha) + (n_\alpha - 1)\psi_\alpha(\hat{X}_\alpha) - \sum_{\beta\subset\alpha} \psi_\alpha(\hat{X}_{\alpha\backslash\beta}, x_\beta) \right] , \end{aligned} \tag{21}$$

*where* $n_\alpha \equiv \sum_{\beta\subset\alpha} 1$.

**Proof** For convenience and without loss of generality, we omit $\alpha$ from our notation and renumber the nodes that are contained in $\alpha$ from 1 to $n$. We consider the quotient between the loopy belief on the potential subset divided by the product of its single-node marginals:

$$\begin{aligned} \frac{Q^*(X)}{\prod\limits_{\beta=1}^{n} Q^*(x_\beta)} &= \frac{\Psi(X)\prod\limits_\beta \mu_\beta(x_\beta) \left[ \sum\limits_{X'} \Psi(X') \prod\limits_\beta \mu_\beta(x'_\beta) \right]^{n-1}}{\prod\limits_\beta \left[ \sum\limits_{X'_{\backslash\beta}} \Psi(X'_{\backslash\beta}, x_\beta) \prod\limits_{\beta'\neq\beta} \mu_{\beta'}(x'_{\beta'}) \mu_\beta(x_\beta) \right]} \\ &= \frac{\Psi(X) \left[ \sum\limits_{X'} \Psi(X') \prod\limits_\beta \mu_\beta(x'_\beta) \right]^{n-1}}{\prod\limits_\beta \left[ \sum\limits_{X'_{\backslash\beta}} \Psi(X'_{\backslash\beta}, x_\beta) \prod\limits_{\beta'\neq\beta} \mu_{\beta'}(x'_{\beta'}) \right]} , \end{aligned} \tag{22}$$

where we substituted the properly normalized version of (5): a loopy belief pseudomarginal is proportional to the potential times incoming messages. The goal is now to find the maximum of the above expression over all possible messages and all values of $X$. Especially the maximum over messages $\mu$ seems to be difficult to compute, but the following intermediate lemma helps us out.

**Lemma 7.2** *The maximum of the function*

$$\begin{aligned} V(\mu) &= (n-1)\log\left[\sum_X \Psi(X)\prod_{\beta=1}^n \mu_\beta(x_\beta)\right] \\ &\quad - \sum_{\beta=1}^n \log\left[\sum_{X_{\setminus\beta}} \Psi(X_{\setminus\beta}, x^*_\beta)\prod_{\beta'\neq\beta}\mu_{\beta'}(x_{\beta'})\right] , \end{aligned}$$

*with respect to the messages $\mu$ under constraints $\sum_{x_\beta}\mu_\beta(x_\beta)$ for all $\beta$ and $\mu_\beta(x_\beta) \geq 0$ for all $\beta$ and $x_\beta$, occurs at an extreme point $\mu_\beta(x_\beta) = \delta_{x_\beta,\hat{x}_\beta}$ for some $\hat{x}_\beta$ to be found.*

**Proof** Let us consider optimizing the message $\mu_1(x_1)$ with fixed messages $\mu_\beta(x_\beta)$ for $\beta > 1$. The first and second derivatives are easily found to obey

$$\begin{aligned} \frac{\partial V}{\partial \mu_1(x_1)} &= (n-1)Q(x_1) - \sum_{\beta\neq 1} Q(x_1|x^*_\beta) \\ \frac{\partial^2 V}{\partial\mu_1(x_1)\partial\mu_1(x'_1)} &= (n-1)Q(x_1)Q(x'_1) - \sum_{\beta\neq 1} Q(x_1|x^*_\beta)Q(x'_1|x^*_\beta) , \end{aligned}$$

where

$$Q(X) \equiv \frac{\Psi(X)\prod_\beta \mu_\beta(x_\beta)}{\sum_{X'}\Psi(X')\prod_\beta\mu_\beta(x'_\beta)} .$$

Now suppose that $V$ has a regular extremum (maximum or minimum) not at an extreme point, i.e., $\mu_1(x_1) > 0$ for two or more values of $x_1$. At such an extremum the first derivative should obey

$$(n-1)Q(x_1) - \sum_{\beta\neq 1} Q(x_1|x^*_\beta) = \lambda ,$$

with $\lambda$ a Lagrange multiplier implementing the constraint $\sum_{x_1}\mu_1(x_1) = 1$. Summing over $x_1$ we obtain $\lambda = 0$ (in fact, $V$ is indifferent to any multiplicative scaling of $\mu$). For the matrix with second derivatives at such an extremum, we then have

$$\frac{\partial^2 V}{\partial\mu_1(x_1)\partial\mu_1(x'_1)} = \sum_{\beta\neq 1}\sum_{\substack{\beta'\neq 1\\ \beta'\neq\beta}} Q(x_1|x^*_\beta)Q(x'_1|x^*_\beta) ,$$

which is positive semidefinite: the extremum cannot be a maximum. Consequently, any maximum must be at the boundary of the domain. Since this holds for any choice of $\mu_\beta(x_\beta)$, $\beta > 1$, it follows by induction that the maximum with respect to all $\mu_\beta(x_\beta)$ must be at an extreme point as well. ∎

The function $V(\mu)$ is, up to a term independent of $\mu$, the logarithm of (22). So, the intermediate Lemma 7.2 tells us that we can replace the maximization over messages $\mu$ by maximization over values $\hat{X}$:

$$\max_{\mu} \frac{Q^*(X)}{\prod_\beta Q^*(x_\beta)} = \max_{\hat{X}} \frac{\Psi(X)\left[\Psi(\hat{X})\right]^{n-1}}{\prod_\beta \Psi(\hat{X}_{\setminus\beta}, x_\beta)} .$$

Next we take the maximum over $X$ as well and define the "strength" $\sigma$ to be used in (20) through

$$\frac{1}{1-\sigma} \equiv \max_{X,\mu} \frac{Q^*(X)}{\prod_\beta Q^*(x_\beta)} = \max_{X,\hat{X}} \frac{\Psi(X)\left[\Psi(\hat{X})\right]^{n-1}}{\prod_\beta \Psi(\hat{X}_{\setminus\beta}, x_\beta)} . \tag{23}$$

The inequality (20) then follows by summing out $X_{\setminus\{\beta,\beta'\}}$ in

$$Q^*(X) - \prod_\beta Q^*(x_\beta) \leq \sigma Q^*(X) .$$

The form (21) then follows by rewriting (23) as

$$\omega \equiv -\log(1-\sigma) = \max_{X,\hat{X}} W(X;\hat{X}) \quad \text{with}$$

$$W(X;\hat{X}) = \left[\psi(X) + (n-1)\psi(\hat{X}) - \sum_\beta \psi(\hat{X}_{\setminus\beta}, x_\beta)\right] ,$$

where we recall that $\psi(X) \equiv \log \Psi(X)$. ∎

## 7.2 Some properties

In the following we will refer to both $\omega$ and $\sigma$ as the strength of the potential. There are several properties worth noting.

- The strength of a potential is indifferent to multiplication with any term that factorizes over the nodes, i.e.,

  $$\text{if } \tilde{\Psi}(X) = \Psi(X)\prod_\beta \mu_\beta(x_\beta) \text{ then } \omega(\tilde{\Psi}) = \omega(\Psi) \text{ for any choice of } \mu.$$

  This property relates to the arbitrariness in the definition of (1): if two potentials overlap, then multiplying one potential with a term that only depends on the overlap and dividing the other by the same term does not change the distribution. Luckily, it also does not change the strength of those potentials.

- To compute the strength, we can enumerate all possible combinations. However, we can neglect all combinations $X$ and $\hat{X}$ that differ in less than two nodes. To see this, consider

$$
\begin{aligned}
&W(x_1, x_2, x_{\backslash 1 \backslash 2}; \hat{x}_1, \hat{x}_2, x_{\backslash 1 \backslash 2}) = \\
&\quad = \psi(x_1, x_2, x_{\backslash 1 \backslash 2}) + \psi(\hat{x}_1, \hat{x}_2, x_{\backslash 1 \backslash 2}) - \psi(\hat{x}_1, x_2, x_{\backslash 1 \backslash 2}) - \psi(x_1, \hat{x}_2, x_{\backslash 1 \backslash 2}) \\
&\quad = -W(x_1, \hat{x}_2, x_{\backslash 1 \backslash 2}; \hat{x}_1, x_2, x_{\backslash 1 \backslash 2}) \, .
\end{aligned}
$$

If now also $\hat{x}_2 = x_2$ we get $W(x_1, x_{\backslash 1}; \hat{x}_1, x_{\backslash 1}) = -W(x_1, x_{\backslash 1}; \hat{x}_1, x_{\backslash 1}) = 0$. Furthermore, if $W(x_1, x_2, x_{\backslash 1 \backslash 2}; \hat{x}_1, \hat{x}_2, x_{\backslash 1 \backslash 2}) \leq 0$ then it must be that $W(x_1, \hat{x}_2, x_{\backslash 1 \backslash 2}; \hat{x}_1, x_2, x_{\backslash 1 \backslash 2}) \geq 0$ and vice versa. So $\omega$, the maximum over all combinations must be non-negative and we can indeed neglect all combinations that by definition yield zero.

- Thus, for finite potentials $0 \leq \omega < \infty$ and $0 \leq \sigma < 1$.

- With pairwise potentials, the above symmetries can be used to reduce the number of evaluations to $|x_1||x_2|(|x_1| - 1)(|x_2| - 1)/4$ combinations. And indeed, for binary nodes $x_{1,2} \in \{0, 1\}$ we immediately obtain

$$
\omega = |\psi(0, 0) + \psi(1, 1) - \psi(0, 1) - \psi(1, 0)| \, . \tag{24}
$$

Any pairwise binary potential can be written as a "Boltzmann factor":

$$
\Psi(x_1, x_2) \propto \exp\left[w x_1 x_2 + \theta_1 x_1 + \theta_2 x_2\right] \, .
$$

In this notation we find the simple and intuitive expression $\omega = |w|$: the strength is the absolute value of the "weight". It is indeed independent of (the size of) the thresholds. In the case of $\{-1, 1\}$ coding the relationship is $\omega = 4|w|$.

- In some models there is the notion of a "temperature" $T$, i.e., $\Psi(X) \propto \exp[\tilde{\psi}(X)/T]$ where $\tilde{\psi}(X)$ is considered constant. In obvious notation we then have $\omega(T) = \omega(1)/T$ and thus $\sigma(T) = 1 - \exp[-\omega(1)/T] = 1 - [1/(1 - \sigma(1))]^{1/T}$.

- Loopy belief revision (max-product) can be interpreted as a zero-temperature limit of loopy belief propagation (sum-product). More specifically, we get the belief revision updates if we imagine running loopy belief propagation on potentials that are scaled with temperature $T$ and then take the limit $T$ to zero. Consequently, when analyzing conditions for uniqueness of loopy belief revision fixed points, we can take $\sigma(0) = 0$ if $\sigma(1) = 0$ (fake interaction), yet $\sigma(0) = 1$ whenever $\sigma(1) > 0$.

# 8 Conditions for uniqueness

## 8.1 Main result

**Theorem 8.1** *Loopy belief propagation has a unique fixed point if there exists an "allocation matrix" $A_{\alpha\beta}$ between potentials $\alpha$ and nodes $\beta$ with properties*

$$
\begin{aligned}
&1.\quad A_{\alpha\beta} \geq 0 \quad \forall_{\alpha,\beta\subset\alpha} && \textit{(positivity)}\\
&2.\quad (1-\sigma_\alpha)\max_{\beta\subset\alpha} A_{\alpha\beta} + \sigma_\alpha \sum_{\beta\subset\alpha} A_{\alpha\beta} \leq 1 \quad \forall_\alpha && \textit{(sufficient amount of resources)}\\
&3.\quad \sum_{\alpha\supset\beta} A_{\alpha\beta} \geq n_\beta - 1 \quad \forall_\beta && \textit{(sufficient compensation)}
\end{aligned}
\tag{25}
$$

*with the strength $\sigma_\alpha$ a function of the potential $\Psi_\alpha(X_\alpha)$ as defined in (21).*

**Proof** For completeness we first summarize our line of reasoning. Fixed points of loopy belief propagation are in one-to-one correspondence with extrema of the dual (16). This dual has a unique extremum if it is "convex/concave". Concavity is guaranteed, so we focus on conditions for convexity, i.e., for positive (semi)definiteness of the corresponding Hessian. This then boils down to conditions that ensure $K = K_1 + \tilde{K}_2 + \tilde{K}_3 \geq 0$ for any choice of $R_\beta(x_\beta)$.

Substituting the bound (20) into the term $\tilde{K}_3$ we obtain

$$
\begin{aligned}
\tilde{K}_3 &\geq -\sum_\alpha \sum_{\substack{\{\beta,\beta'\}\subset\alpha \\ \beta'\neq\beta}} \sum_{x_\beta, x'_{\beta'}} A_{\alpha\beta} A_{\alpha\beta'} \sigma_\alpha Q^*_\alpha(x_\beta, x'_{\beta'}) R_\beta(x_\beta) R_{\beta'}(x'_{\beta'}) \\
&\geq -\sum_\alpha \sigma_\alpha \sum_{\beta\subset\alpha} \sum_{x_\beta} A_{\alpha\beta} \sum_{\substack{\beta'\subset\alpha \\ \beta'\neq\beta}} A_{\alpha\beta'} Q^*_\alpha(x_\beta) R^2_\beta(x_\beta) \,,
\end{aligned}
$$

where in the last step we applied the same trick as in (19). Since $\tilde{K}_2 \geq 0$ and combining $K_1$ and (the above lower bound on) $\tilde{K}_3$, we get

$$
K = K_1 + \tilde{K}_2 + \tilde{K}_3 \geq \sum_\alpha \sum_{\beta\subset\alpha} \sum_{x_\beta} A_{\alpha\beta} \left[ 1 - A_{\alpha\beta} - \sigma_\alpha \sum_{\beta'\neq\beta} A_{\alpha\beta'} \right] Q^*_\alpha(x_\beta) R^2_\beta(x_\beta) \,.
$$

This implies

$$
(1-\sigma_\alpha) A_{\alpha\beta} + \sigma_\alpha \sum_{\beta'\subset\alpha} A_{\alpha\beta'} \leq 1 \quad \forall_{\alpha,\beta\subset\alpha} \,,
$$

which, in combination with $A_{\alpha\beta} \geq 0$ and $\sigma_\alpha \leq 1$, yields condition 2. in (25). The equality constraint (13) that we started with can be relaxed to the inequality condition 3. without any consequences. ∎

We get back the more strict conditions of Theorem 4.2 if $\sigma_\alpha = 1$ for all potentials $\alpha$. Furthermore, "fake interactions" play no role: with $\sigma_\alpha = 0$ condition 2. becomes $\max_{\beta\subset\alpha} A_{\alpha\beta} \leq 1$ suggesting the choice $A_{\alpha\beta} = 1$ for all $\beta \subset \alpha$, which then effectively reduces the number of neighboring potentials $n_\beta$ in condition 3.

## 8.2 Comparison with other work

To the best of our knowledge, the only conditions for uniqueness of loopy belief propagation fixed points that depend on more than just the structure of the graph are those in (Tatikonda & Jordan 2002) for pairwise potentials. The analysis in (Tatikonda & Jordan 2002) is based on the concept of the computation tree, which represents an unwrapping of the original graph with respect to the loopy belief propagation algorithm. The same concept is used in (Weiss 2000) to show that belief revision yields the correct maximum a posteriori assignments in graphs with a single loop and (Weiss &Freeman 2001) to prove that loopy belief propagation in Gaussian graphical models yields exact means. Although the current theorems based on the concept of computation trees are derived for pairwise potentials, it should be possible to extend them to more general factor graphs.

The setup in (Tatikonda & Jordan 2002) is slightly different, namely based on the factorization

$$P_{\text{exact}}(X) = \frac{1}{Z} \prod_{\alpha} \hat{\Psi}_{\alpha}(X_{\alpha}) \prod_{\beta} \hat{\Psi}_{\beta}(x_{\beta}) \; ,$$

to be compared with our (1) where there are no self-potentials $\Psi_{\beta}(x_{\beta})$. With this in mind, the statement is then as follows.

**Theorem 8.2** *Adapted from (Tatikonda & Jordan 2002), in particular Proposition 5.3. Loopy belief propagation on pairwise potentials has a unique fixed point if*

$$\sum_{\alpha \supset \beta} \left( \max_{X_\alpha} \hat{\psi}_{\alpha}(X_{\alpha}) - \min_{X_\alpha} \hat{\psi}_{\alpha}(X_{\alpha}) \right) < 2 \quad \forall_{\beta} \; , \tag{26}$$

To make the connection between Theorem 8.2 and Theorem 8.1, we will first strengthen the former and then weaken the latter. We will focus on the case of binary pairwise potentials. Since the definition of self-potentials is arbitrary and the condition (26) is valid for any choice, we can easily improve the condition by optimizing this choice. This then leads to the following corollary.

**Corollary 8.3** *Improvement of Theorem 8.2 for pairwise binary potentials. Loopy belief propagation on pairwise binary potentials has a unique fixed point if*

$$\sum_{\alpha \supset \beta} \omega_{\alpha} < 4 \quad \forall_{\beta} \; , \tag{27}$$

*with $\omega_\alpha$ defined in (21).*

**Proof** The condition (26) applies to any arbitrary definition of self-potentials $\hat{\Psi}_\beta(x_\beta)$. In fact, it is valid for any choice

$$\hat{\psi}_\alpha(X_\alpha) = \psi_\alpha(X_\alpha) + \sum_{\beta \subset \alpha} \phi_{\alpha\beta}(x_\beta) \ ,$$

where $\psi_\alpha(X_\alpha)$ is any choice of potential subsets that fits in our framework of no self-potentials (as argued above, there is some arbitrariness here as well). We can then optimize this choice to obtain milder and thus better conditions. Omitting $\alpha$ and renumbering the nodes from 1 to 2, we have

$$\min_{\phi_1,\phi_2} \left\{ \max_{x_1,x_2} \hat{\psi}(x_1,x_2) - \min_{x_1,x_2} \hat{\psi}(x_1,x_2) \right\} =$$
$$\min_{\phi_1,\phi_2} \left\{ \max_{x_1,x_2} \left[\psi(x_1,x_2) + \phi_1(x_1) + \phi_2(x_2)\right] - \min_{x_1,x_2} \left[\psi(x_1,x_2) + \phi_1(x_1) + \phi_2(x_2)\right] \right\} \ .$$

In the case of binary nodes (two-by-two matrices $\psi(x_1,x_2)$), it is easy to check that the optimal $\phi_1$ and $\phi_2$ that yield the smallest gap are such that

$$\begin{aligned} \psi(x_1,x_2) + \phi_1(x_1) + \phi_2(x_2) = \psi(\hat{x}_1,\hat{x}_2) + \phi_1(\hat{x}_1) + \phi_2(\hat{x}_2) \quad \geq \\ \psi(x_1,\hat{x}_2) + \phi_1(x_1) + \phi_2(\hat{x}_2) = \psi(\hat{x}_1,x_2) + \phi_1(\hat{x}_1) + \phi_2(x_2) \ , \end{aligned} \tag{28}$$

for some $x_1, x_2, \hat{x}_1$ and $\hat{x}_2$ with $x_1 \neq \hat{x}_1$ and $x_2 \neq \hat{x}_2$. Solving for $\phi_1$ and $\phi_2$ we find

$$\begin{aligned} \phi_1(x_1) - \phi_1(\hat{x}_1) &= \frac{1}{2}\left[\psi(\hat{x}_1,x_2) - \psi(x_1,\hat{x}_2) + \psi(\hat{x}_1,\hat{x}_2) - \psi(x_1,x_2)\right] \\ \phi_2(x_2) - \phi_2(\hat{x}_2) &= \frac{1}{2}\left[\psi(x_1,\hat{x}_2) - \psi(\hat{x}_1,\hat{x}_2) + \psi(\hat{x}_1,\hat{x}_2) - \psi(x_1,x_2)\right] \ . \end{aligned}$$

Substitution back into (28) yields

$$\begin{aligned} \psi(x_1,x_2) + \phi_1(x_1) + \phi_2(x_2) - \psi(x_1,\hat{x}_2) - \phi_1(x_1) - \phi_2(\hat{x}_2) \ = \\ \frac{1}{2}\left[\psi(x_1,x_2) + \psi(\hat{x}_1,\hat{x}_2) - \psi(\hat{x}_1,x_2) - \psi(x_1,\hat{x}_2)\right] \ , \end{aligned}$$

which has to be non-negative. Off all 4 possible combinations, two of them are valid and yield the same positive gap, and the other two are invalid since they yield the same negative gap. Enumerating these combinations, we find

$$\begin{aligned} \min_{\phi_1,\phi_2} \left\{ \max_{x_1,x_2} \hat{\psi}(x_1,x_2) - \min_{x_1,x_2} \hat{\psi}(x_1,x_2) \right\} &= \frac{1}{2}|\psi(0,0) + \psi(1,1) - \psi(0,1) - \psi(1,0)| \\ &= \frac{\omega}{2} \ , \end{aligned}$$

from (24). Substitution into the condition (26) then yields (27). ∎

Next we derive the following weaker corollary of Theorem 8.1.

**Corollary 8.4** *Weaker version of Theorem 8.1 for pairwise potentials. Loopy belief propagation on pairwise potentials has a unique fixed point if*

$$\sum_{\alpha \supset \beta} \omega_\alpha \leq 1 \quad \forall_\beta \,, \tag{29}$$

*with $\omega_\alpha$ defined in (21).*

**Proof** Consider the allocation matrix with components $A_{\alpha\beta} = 1 - \sigma_\alpha$ for all $\beta \subset \alpha$. With this choice, conditions 1. and 2. of (25) are fulfilled, since (condition 1.) $\sigma_\alpha \leq 1$ and (condition 2.)

$$(1-\sigma_\alpha)(1-\sigma_\alpha) + 2\sigma_\alpha(1-\sigma_\alpha) = 1 - 2\sigma_\alpha^2 \leq 1 \,.$$

Substitution into condition 3. yields

$$\sum_{\alpha \supset \beta} (1-\sigma_\alpha) \geq \sum_{\alpha \supset \beta} 1 - 1 \quad \text{and thus} \quad \sum_{\alpha \supset \beta} \sigma_\alpha \leq 1 \,. \tag{30}$$

Since $\omega_\alpha = -\log(1-\sigma_\alpha) \geq \sigma_\alpha$, condition (29) is weaker than condition (30). ∎

Summarizing, the conditions in (Tatikonda & Jordan 2002) are, for binary pairwise potentials and when strengthened as above, at most a constant (factor 4) less strict and thus better than the ones derived here. The latter are better when the structure is (close to) a tree. The best set of conditions follows by taking the union of both. Note further that the conditions derived in (Tatikonda & Jordan 2002) are, unlike Theorem 8.1, specific to pairwise potentials.

## 8.3 Illustration

For illustration we consider a $3 \times 3$ Ising grid with toroidal boundary conditions as in Figure 3(a) and uniform ferromagnetic potentials proportional to

$$\begin{pmatrix} \alpha & 1-\alpha \\ 1-\alpha & \alpha \end{pmatrix} .$$

The trivial solution, which is the only minimum of the Bethe free energy for small $\alpha$, is the one with all pseudomarginals equal to $(0.5, 0.5)$. With simple algebra, e.g., following the line of reasoning that leads to the belief optimization algorithm in (Welling &Teh 2003), it can be shown that this trivial solution becomes unstable at the critical $\alpha_{\text{critical}} = 2/3 \approx 0.67$. For $\alpha > 2/3$ we find two minima: one with "spins up" and the other one with "spins down".

In this symmetric problem, the strength of each potential is given by

$$\omega = 2 \log \left[ \frac{\alpha}{1-\alpha} \right] \quad \text{and thus} \quad \sigma = 1 - \left( \frac{1-\alpha}{\alpha} \right)^2 .$$

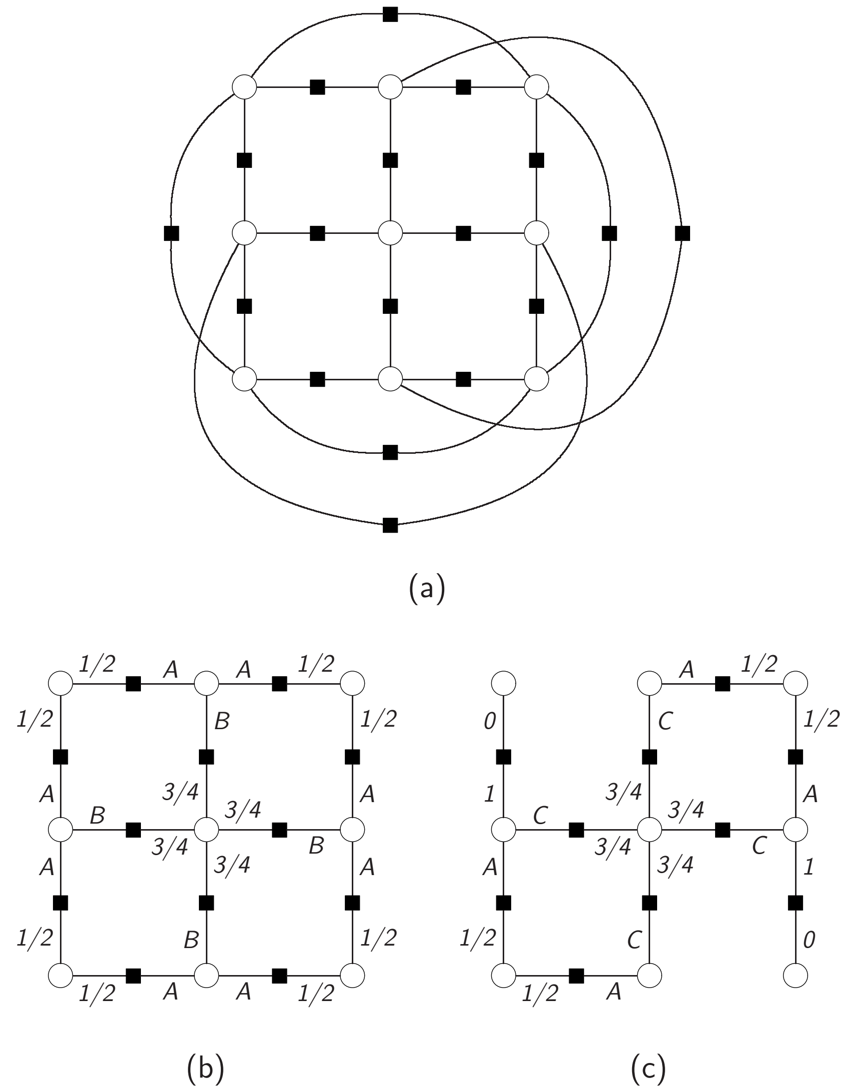



Figure 3: Three Ising grids in factor-graph notation: circles denote nodes, boxes interactions. (a) With toroidal boundary conditions. All elements of the allocation matrix equal to 3/4 (not shown). (b) With aperiodic boundary conditions and (c) two loops left. The elements of the allocation matrix along the edges follow directly from optimizing condition 3. in Theorem 8.1 and symmetry considerations. With $B = 2 - 2A$ in (b) and $C = 1 - A$ in (c), the optimal settings for the single remaining variable $A$ then boil down to 3/4 and $1 - \sqrt{1/8}$, respectively. See the main text for further explanation.

The minimal (uniform) compensation in condition 3. of Theorem 8.1 amounts to $A = 3/4$ for all combinations of potentials and nodes. Substitution into condition 2. then yields

$$\sigma \leq \frac{1}{3} \text{ and thus } \alpha \leq \frac{1}{1+\sqrt{2/3}} \approx 0.55 \,.$$

The critical value that follows from Corollary 8.3 is in this case slightly better:

$$\omega < 1 \text{ and thus } \alpha \leq \frac{1}{1+\mathrm{e}^{-1/2}} \approx 0.62 \,.$$

Next we consider the same grid with aperiodic boundary conditions as in Figure 3(b). Numerically, we find a critical $\alpha_{\text{critical}} \approx 0.79$. The value that follows from Corollary 8.3 is dominated by the center node and hence stays the same: a unique loopy belief propagation fixed point for $\alpha < 0.62$. Theorem 8.1 on the other hand can be exploited to shift resources a little. In principle we can solve the nonlinear programming problem, but for this small problem it can still be done by hand with the following argumentation. Minimal compensation according to condition 3. in Theorem 8.1 combined with symmetry considerations yields the allocation matrix elements along on the edges in Figure 3(b). It is then easy to check that there are only two different appearances of condition 2., namely

$$(2-2A)\sigma + \frac{3}{4} \leq 1 \text{ and } \frac{1}{2}\sigma + A \leq 1 \,.$$

The optimal choice for $A$ is the one in which both conditions turn out the identical. In this way we obtain $A = 3/4$, yielding

$$\sigma \leq \frac{1}{2} \text{ and thus } \alpha \leq \frac{1}{1+\sqrt{1/2}} \approx 0.58 \,,$$

still slightly worse then the condition from Corollary 8.3.

An example in which the condition obtained with Theorem 8.1 is better than the one from Corollary 8.3 is given in Figure 3(c). Straightforward analysis following the same recipe as for Figure 3(b) yields $A = 1 - \sqrt{1/8}$ with

$$\sigma \leq \sqrt{\frac{1}{2}} \text{ and thus } \alpha \leq \frac{1}{1+\sqrt{1-\sqrt{1/2}}} \approx 0.65 \,,$$

better than the $\alpha < 0.62$ from Corollary 8.3 and to be compared with the critical $\alpha_{\text{critical}} \approx 0.88$.

# 9 Discussion

In this article, we derived sufficient conditions for loopy belief propagation to have just a single fixed point. It is for sure that these conditions are still much too strong to be anywhere near the necessary conditions and in that sense should be seen as no more than a first step. Nice features of these conditions are that they

- generalize the conditions for convexity of the Bethe free energy;
- incorporate the (local) strength of potentials;
- scale naturally as a function of the "temperature";
- are invariant to arbitrary definitions of potentials and self-interactions.

Although the analysis that led to these conditions may seem quite involved, it basically consists of a relatively straightforward combination of two observations. First, that we can exploit the arbitrariness in the definition of the Bethe free energy when we incorporate the constraints. This forms the basis of the resource allocation argument. And second, the bound on the correlation of a loopy belief propagation marginal that leads to the introduction of the strength of a potential.

Besides its theoretical usefulness, we can think of the following more practical uses.

**Convergent algorithms.** Algorithms for guaranteed convergence explicitly minimize the Bethe free energy. They can be considered "bound optimization algorithms", similar to expectation maximization and iterative proportional fitting: in the inner loop they minimize a bound on the Bethe free energy, which is then updated in the outer loop. In practice it appears that the tighter the bound, the faster the convergence (see e.g. (Heskes et al. 2003)). Instead of a bound that is convex (Yuille 2002) or convex over the set of constraints (Teh & Welling 2002, Heskes et al. 2003), we might relax the convexity condition and choose a tighter bound that still has a unique minimum, thereby speeding up the convergence.

**Other free energies.** In (Wainwright et al. 2002) a convexified Bethe free energy is proposed. The arguments for this class of free energies are two-fold: they yield a bound on the partition function (instead of just an approximation, as the standard Bethe free energy) and they have a unique minimum. Focusing on the second argument, the conditions in this article can be used to construct Bethe free energies that may not be convex (over the set of constraints), but do have a unique minimum and, being closer to the standard Bethe free energy, may yield better approximations.

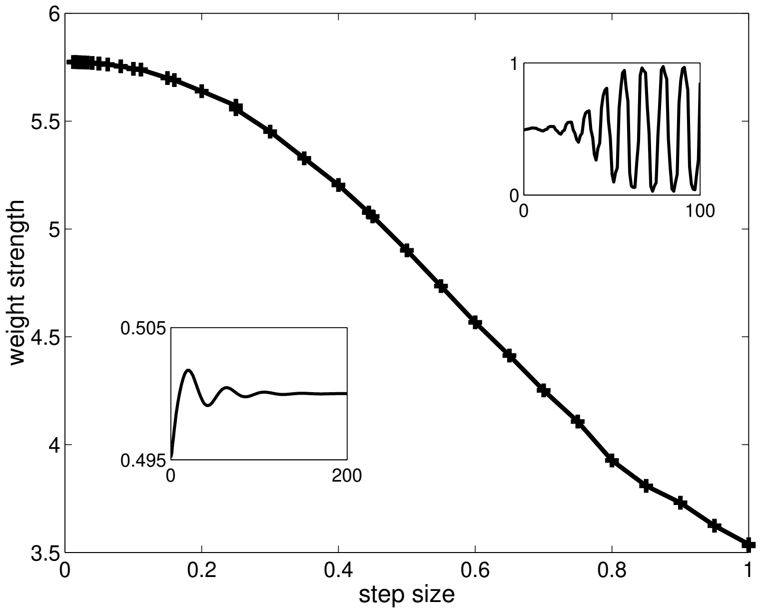



Figure 4: The transition between "convergent" and "non-convergent" behavior as a function of the step size used for damping loopy belief propagation and the weight strength. Simulations on a 4-node Boltzmann machine. The insets show the marginal $P_1(x_1 = 1)$ as a function of the number of loopy belief iterations for step size 0.2 and strength 4 (lower left) and step size 0.6 and strength 6 (upper right). See text for further details.

We can think of the following opportunities to make the sufficient conditions derived here more strict and thus closer to necessary conditions.

- The conditions guarantee convexity of the dual $G(Q_\beta, \lambda_{\alpha\beta})$ with respect to $Q_\beta$. But in fact we need only $G(Q_\beta) \equiv \max_{\lambda_{\alpha\beta}} G(Q_\beta, \lambda_{\alpha\beta})$ to be convex, which is a weaker requirement. The Hessian of $G(Q_\beta)$, however, appears to be more difficult to compute and to analyze in general, but may lead to stronger results in specific cases (e.g., only pairwise interactions or substituting a particular choice of $A_{\alpha\beta}$).

- It may be possible to strengthen the bound (20) on loopy belief correlations, especially for interactions that involve more than two nodes.

An important question is how the uniqueness of loopy belief propagation fixed points relates to the convergence of loopy belief propagation. Intuitively one might expect that if loopy belief propagation has a unique fixed point, it will also converge to it. This also seems to be the argumentation in (Tatikonda & Jordan 2002). However, to the best of our knowledge there is no proof of such correspondence. Furthermore, the following set of simulations does seem to suggest otherwise.

We consider a Boltzmann machine with 4 binary nodes, weights

$$w = \omega \begin{pmatrix} 0 & 1 & -1 & -1 \\ 1 & 0 & 1 & -1 \\ -1 & 1 & 0 & -1 \\ -1 & -1 & -1 & 0 \end{pmatrix},$$

zero thresholds, and potentials

$$\Psi_{ij}(x_i, x_j) = \exp\left[w_{ij}/4\right] \text{ if } x_i = x_j \quad \text{and} \quad \Psi_{ij}(x_i, x_j) = \exp\left[-w_{ij}/4\right] \text{ if } x_i \neq x_j \ .$$

Running loopy belief propagation, possibly damped as in (9), we observe "convergent" and "non-convergent" behavior. For relatively small weights, loopy belief propagation converges to the trivial fixed point with $P_i(x_i) = 0.5$ for all nodes $i$ and $x_i = \{0, 1\}$, as in the lower left inset in Figure 4. For relatively large weights, it ends up in a limit cycle yield exactly the same condition for $\sigma$: as shown in the upper right inset. The weight strength that forms the transition between this "convergent" and "non-convergent" behavior strongly depends on the step size[7]. This by itself makes it hard to defend a one-to-one correspondence between convergence of loopy belief propagation (apparently depending on step size) and uniqueness of fixed points (obviously independent of step size).

[7]Note that the conditions for guaranteed uniqueness imply $\omega = 4/3$ for Corollary 8.3 and $\omega = \log(2) \approx 0.69$ for Theorem 8.1, both far below the weight strengths where "non-convergent" behavior sets in.

For weights larger than roughly 5.8, loopy belief propagation failed to converge to the trivial fixed point even for very small step sizes. However, running a convergent double-loop algorithm from many different initial conditions and many weight strengths considerably larger than 5.8, we always ended up in the trivial fixed point and never in another one. We found similar behavior for a 3-node Boltzmann machine (same weight matrix as above, except for the fourth node) for very large weights: loopy belief propagation ends up in a limit cycle, whereas a convergent double-loop algorithm converges to the trivial fixed point, which here, by Corollary 4.4, is guaranteed to be unique. In future work we hope to elaborate on these issues.


## Acknowledgements

This work has been supported in part by the Dutch Technology Foundation STW. I would like to thank the anonymous reviewers for their constructive comments and Joris Mooij for computing the critical $\alpha_{\text{critical}}$'s in Section 8.3.


# References


Heskes, T. (2002), Stable fixed points of loopy belief propagation are minima of the Bethe free energy, *in* S. Becker, S. Thrun & K. Obermayer, eds, 'Advances in Neural Information Processing Systems 15', MIT Press, Cambridge, pp. 359–366.

Heskes, T., Albers, K. & Kappen, B. (2003), Approximate inference and constrained optimization, *in* 'Uncertainty in Artificial Intelligence: Proceedings of the Nineteenth Conference (UAI-2003)', Morgan Kaufmann Publishers, San Francisco, CA, pp. 313–320.

Kschischang, F., Frey, B. & Loeliger, H. (2001), 'Factor graphs and the sum-product algorithm', *IEEE Transactions on Information Theory* **47**(2), 498–519.

Lauritzen, S. & Spiegelhalter, D. (1988), 'Local computations with probabilities on graphical structures and their application to expert systems', *Journal of the Royal Statistics Society B* **50**, 157–224.

Luenberger, D. (1984), *Linear and Nonlinear Programming*, Addison-Wesley, Reading, Massachusetts.

McEliece, R., MacKay, D. & Cheng, J. (1998), 'Turbo decoding as as an instance of Pearl's 'belief propagation' algorithm', *IEEE Journal on Selected Areas in Communication* **16**(2), 140–152.

McEliece, R. & Yildirim, M. (2003), Belief propagation on partially ordered sets, *in* D. Gilliam & J. Rosenthal, eds, ‘Mathematical Systems Theory in Biology, Communications, Computation, and Finance’, Springer, New York, pp. 275–300.

Minka, T. (2001), Expectation propagation for approximate Bayesian inference, *in* J. Breese & D. Koller, eds, ‘Uncertainty in Artificial Intelligence: Proceedings of the Seventeenth Conference (UAI-2001)’, Morgan Kaufmann Publishers, San Francisco, CA, pp. 362–369.

Murphy, K., Weiss, Y. & Jordan, M. (1999), Loopy belief propagation for approximate inference: An empirical study, *in* K. Laskey & H. Prade, eds, ‘Proceedings of the Fifteenth Conference on Uncertainty in Articial Intelligence’, Morgan Kaufmann Publishers, San Francisco, CA, pp. 467–475.

Pakzad, P. & Anantharam, V. (2002), Belief propagation and statistical physics, *in* ‘2002 Conference on Information Sciences and Systems’, Princeton University.

Pearl, J. (1988), *Probabilistic Reasoning in Intelligent systems: Networks of Plausible Inference*, Morgan Kaufmann, San Francisco, CA.

Tatikonda, S. & Jordan, M. (2002), Loopy belief propagation and Gibbs measures, *in* A. Darwiche & N. Friedman, eds, ‘Uncertainty in Artificial Intelligence: Proceedings of the Eighteenth Conference (UAI-2002)’, Morgan Kaufmann Publishers, San Francisco, CA, pp. 493–500.

Teh, Y. & Welling, M. (2002), The unified propagation and scaling algorithm, *in* T. Dietterich, S. Becker & Z. Ghahramani, eds, ‘Advances in Neural Information Processing Systems 14’, MIT Press, Cambridge, pp. 953–960.

Wainwright, M., Jaakkola, T. & Willsky, A. (2002), A new class of upper bounds on the log partition function, *in* A. Darwiche & N. Friedman, eds, ‘Uncertainty in Artificial Intelligence: Proceedings of the Eighteenth Conference (UAI-2002)’, Morgan Kaufmann Publishers, San Francisco, CA, pp. 536–543.

Weiss, Y. (2000), ‘Correctness of local probability propagation in graphical models with loops’, *Neural Computation* **12**(1), 1–41.

Weiss, Y. &Freeman, W. (2001), ‘Correctness of belief propagation in graphical models with arbitrary topology’, *Neural Computation* **13**(10), 2173–2200.

Welling, M. &Teh, Y. (2003), ‘Approximate inference in Boltzmann machines’, *Artificial Intelligence* **143**(1), 19–50.

Yedidia, J., Freeman, W. & Weiss, Y. (2001), Generalized belief propagation, *in* T. Leen, T. Dietterich & V. Tresp, eds, 'Advances in Neural Information Processing Systems 13', MIT Press, Cambridge, pp. 689–695.

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