Journal of Artificial Intelligence Research 26 (2006) 153-190 Submitted 10/05; published 06/06

Convexity Arguments for Efficient Minimization of the Bethe and Kikuchi Free Energies

Tom Heskes T.HESKES@SCIENCE.RU.NL
IRIS, Faculty of Science, Radboud University Nijmegen
Toernooiveld 1, 6525 ED, Nijmegen, The Netherlands

Abstract

Loopy and generalized belief propagation are popular algorithms for approximate inference in Markov random fields and Bayesian networks. Fixed points of these algorithms have been shown to correspond to extrema of the Bethe and Kikuchi free energy, both of which are approximations of the exact Helmholtz free energy. However, belief propagation does not always converge, which motivates approaches that explicitly minimize the Kikuchi/Bethe free energy, such as CCCP and UPS.

Here we describe a class of algorithms that solves this typically *non-convex* constrained minimization problem through a sequence of *convex* constrained minimizations of upper bounds on the Kikuchi free energy. Intuitively one would expect tighter bounds to lead to faster algorithms, which is indeed convincingly demonstrated in our simulations. Several ideas are applied to obtain tight convex bounds that yield dramatic speed-ups over CCCP.

1. Introduction

Pearl's belief propagation (Pearl, 1988) is a popular algorithm for inference in Bayesian networks. It is known to be exact in special cases, e.g., for tree-structured (singly connected) networks with just Gaussian or just discrete nodes. But also on networks containing cycles, so-called loopy belief propagation empirically often leads to good performance (approximate marginals close to exact marginals) (Murphy, Weiss, & Jordan, 1999; McEliece, MacKay, & Cheng, 1998). 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.

The Kikuchi free energy (Kikuchi, 1951) is a generalization of the Bethe free energy that can lead to better approximations of the exact Helmholtz free energy. Just like fixed points of loopy belief propagation correspond to extrema of the Bethe free energy, fixed points of an algorithm called generalized belief propagation (Yedidia et al., 2001) correspond to extrema of the Kikuchi free energy.

A problem with loopy and generalized belief propagation is that they do not always converge to a stable fixed point. New algorithms (Yuille, 2002; Teh & Welling, 2002) have been derived that therefore explicitly minimize the Bethe and Kikuchi free energy. As we will describe in Section 2, minimization of the Kikuchi free energy corresponds to a usually non-convex constrained minimization problem. Non-convex constrained minimization problems are known to be rather difficult to solve, so in Section 3 we will first derive sufficient conditions for the Kikuchi free energy to be convex (over the set of constraints). In Section 4 we will then derive a class of converging double-loop algorithms, in which each inner loop corresponds to constrained minimization of a *convex bound* on the Kikuchi free energy, and each outer-loop step to a recalculation of this bound. Based on the intuition that the tightest bound yields the fastest algorithm, we come up with several ideas to construct tight bounds. We will see that Yuille's (2002) CCCP algorithm corresponds to a special case of a rather loose bound and discuss the relationship with the UPS algorithm by Teh and Welling (2002) in Section 4.5. The simulations in Section 5 illustrate the use of tight convex bounds on several inference problems. Implications and other issues are discussed in Section 6. Technical details are treated in the appendices.

2. The Kikuchi Approximation

Exact inference in graphical models is often intractable. In this section we will introduce the Kikuchi approximation as a particular example of a variational approach towards approximate inference.

2.1 Graphical Models

An undirected graph $G = (V, E)$ consists of set of nodes or vertices $V = \{1, \ldots, N\}$ that are joined by a set of edges E. We place at each node i a variable x_i which takes values in a finite discrete alphabet. The vector containing all variables is denoted $\mathbf{x} \equiv (x_1, \ldots, x_n)$. Let γ be a subset of V; we call γ a *region*. A *clique* is any fully connected subset of V; $\mathcal{C}$ is a set of cliques. The *potential*, also referred to as compatibility or kernel function, $\psi_\alpha(\mathbf{x}_\alpha)$ is a strictly positive function that only depends on the variables that are part of the clique α. We define the probability distribution or probability mass function

$$p_{\text{exact}}(\mathbf{x}) \equiv \frac{1}{Z} \prod_{\alpha \in \mathcal{C}} \psi_\alpha(\mathbf{x}_\alpha) \,, \tag{1}$$

where Z is the normalizing constant, often called *partition function*. The Hammersley-Clifford theorem (Besag, 1974) guarantees us that the underlying probability process is Markov with respect to the graph and, vice versa, that the distribution of any Markov random field over G that is strictly positive can be expressed in this form. Through the process of moralization, any directed graphical model (Bayesian network) can be transformed into a corresponding undirected model. Consequently, the probability distribution corresponding to a Bayesian network can also be written in the form (1) (Lauritzen, 1996).

Computing the partition function Z, as well as computing marginals on subsets of variables, in principle requires summation over an exponential number of states. To circumvent this exponential summation there are two kinds of approaches: sampling techniques and variational methods. With sampling, one draws samples from the exact probability distribution. The variational methods try to find an approximation to the exact probability distribution.

2.2 Variational Methods

Variational methods are often derived from an approximation of the so-called free energy

$$F(p) = -\sum_{\alpha \in \mathcal{C}} \sum_{\mathbf{x}_\alpha} p(\mathbf{x}_\alpha) \log \psi_\alpha(\mathbf{x}_\alpha) + \sum_{\mathbf{x}} p(\mathbf{x}) \log p(\mathbf{x}) \equiv E(p) - S(p) \,. \tag{2}$$

The first term, $E(p)$, is referred to as the *energy*, the second term $S(p)$ as the *entropy*. Functional minimization of $F(p)$ with respect to functions $p(\mathbf{x})$ under the constraint that $p(\mathbf{x})$ is properly normalized yields $p_{\text{exact}}(\mathbf{x})$. Furthermore, the partition function Z then follows from

$$-\log Z = F(p_{\text{exact}}) \, .$$

When we stick to the exact free energy (2), we do not really gain anything: the entropy part $S(p)$ still consists of a sum over exponentially many terms. Variational methods are based on a tractable approximation of the free energy. They can be roughly divided into two classes, the "mean-field" and the "Kikuchi" approximations. In the mean-field approach one confines the minimization of the free energy to a restricted class $\mathcal{T}$ of (tractable) probability distributions instead of considering the class $\mathcal{P}$ of all probability distributions:

$$-\log Z = F(p_{\text{exact}}) = \min_{p \in \mathcal{P}} F(p) \leq \min_{p \in \mathcal{T}} F(p) \, .$$

The crux is to choose the class $\mathcal{T}$ such that the entropy $S(p)$ becomes tractable for all $p \in \mathcal{T}$. Note however that this restriction typically also affects the energy term $E(p)$ (Jordan, Ghahramani, Jaakkola, & Saul, 1998; Jaakkola & Jordan, 1999).

The Kikuchi approximation of the free energy (2) leaves the energy term as is and approximates the entropy $S(p)$ through a combination of marginal entropies:

$$\begin{aligned} -S(p) &= \sum_{\mathbf{x}} p(\mathbf{x}) \log p(\mathbf{x}) \approx -\sum_{\gamma \in \mathcal{R}} c_\gamma S_\gamma(p) \\ &= \sum_{\gamma \in \mathcal{R}} c_\gamma \sum_{\mathbf{x}_\gamma} p(\mathbf{x}_\gamma) \log p(\mathbf{x}_\gamma) \, . \end{aligned} \tag{3}$$

Here $\mathcal{R}$ denotes a collection of so-called *regions*; the parameters c_γ are called *Moebius* or *overcounting numbers*.

2.2.1 Partially Ordered Sets

Following Pakzad and Anantharam (2002, 2005), we will use the language of *partially ordered sets* or *posets*. Specifically, the collection $\mathcal{R}$ of regions can be viewed as such a poset where the ordering is defined with respect to the inclusion operator $\subset$. A region γ includes a region γ', written $\gamma \supseteq \gamma'$, if all variables in γ' are also part of γ. We use $\gamma \supset \gamma'$ to denote strict inclusion, i.e., $\gamma \supseteq \gamma'$ and $\gamma' \not\supseteq \gamma$. We say that γ covers γ'' in $\mathcal{R}$, written $\gamma \succ \gamma''$, if $\gamma \supset \gamma''$ and there exists no $\gamma' \in \mathcal{R}$ such that $\gamma \supset \gamma' \supset \gamma'$. We can visualize a poset with a so-called *Hasse diagram* or *region graph* (see the examples below). Given a particular poset $\mathcal{R}$, its Hasse diagram $G_{\mathcal{R}}$ is a directed acyclic graph, whose vertices are the elements of $\mathcal{R}$, and whose edges corresponds to the cover relationships. That is, there is an edge from γ to γ' iff $\gamma \succ \gamma'$.

2.3 The Cluster Variation Method

In Kikuchi's (1951) original cluster variation method (CVM), the collections of regions and overcounting numbers are constructed as follows. We start by defining a collection $\mathcal{O}$ of *outer* regions. The minimal choice is the original set of cliques $\mathcal{C}$, but we can also choose to combine cliques and construct larger ones, similar to the process of triangulation (Lauritzen, 1996). For convenience, we redefine the potentials correspondingly, i.e., such that there is precisely one potential $\psi_\alpha(\mathbf{x}_\alpha)$ per outer region α (see the example below).

Given these outer regions, we construct new regions by taking the intersections of these outer regions, the intersections of intersections, and so on, until no more intersections can be made. We will refer to the regions constructed in this way as *inner* regions, combined in the collection $\mathcal{I}$. The collection of all regions $\mathcal{R}$ in (3) is now the union of the outer and inner regions: $\mathcal{R} = \mathcal{O} \cup \mathcal{I}$.

The overcounting or Moebius numbers in the original CVM follow from the Moebius formula

$$c_\gamma = 1 - \sum_{\gamma' \supset \gamma} c_{\gamma'} \,. \tag{4}$$

By definition we have $c_\alpha = 1$ for all outer regions $\alpha \in \mathcal{O}$.

The Bethe free energy can be considered a special case of the Kikuchi free energy. In the Bethe free energy there are no intersections of intersections, i.e., there is only one level of inner regions with $c_\beta = 1 - n_\beta$ where $n_\beta \equiv \sum_{\alpha \in \mathcal{O}; \alpha \supset \beta} 1$ equals the number of outer regions covering inner region β.

2.3.1 Alternatives

Several alternatives to the original CVM, with weaker constraints and/or other constraints on the choice of regions and overcounting numbers, have been proposed recently. Yedidia, Freeman, and Weiss (2005) present an overview. The particular choice of inner regions subsets and overcounting numbers in junction graphs (Aji & McEliece, 2001) and join graphs (Dechter, Kask, & Mateescu, 2002) leads to an entropy approximation in which all overcounting numbers for the inner regions are negative. The resulting algorithms are very similar to the junction tree algorithm, but then applied to a graph with loops. The entropy approximation that follows from the original cluster variation method takes into account all entropy contributions up to the level of the outer regions in a consistent manner and, on theoretical grounds, there seems to be no reason to deviate from that (Pakzad & Anantharam, 2005). In this paper, we therefore focus on the original cluster variation method, but our analysis holds much more generally for any poset or region graph.

2.4 Constrained Minimization

The Kikuchi approximation of the free energy only depends on the marginals $p(\mathbf{x}_\gamma)$ for all $\gamma \in \mathcal{R}$. We now replace the minimization of the exact free energy over the complete distribution $p(\mathbf{x})$ by minimization of the Kikuchi free energy

$$F_{\text{Kikuchi}}(\mathbf{q}) = -\sum_{\alpha \in \mathcal{O}} \sum_{\mathbf{x}_\alpha} q_\alpha(\mathbf{x}_\alpha) \log \psi_\alpha(\mathbf{x}_\alpha) + \sum_{\gamma \in \mathcal{R}} c_\gamma \sum_{\mathbf{x}_\gamma} q_\gamma(\mathbf{x}_\gamma) \log q_\gamma(\mathbf{x}_\gamma) \tag{5}$$

over all *pseudo-marginals* $\mathbf{q} \equiv \{q_\gamma; \gamma \in \mathcal{R}\}$ under the consistency and normalization constraints

$$q_\gamma(\mathbf{x}_\gamma) \geq 0 \quad \forall_{\gamma \in \mathcal{R}} \forall_{\mathbf{x}_\gamma} \qquad \textit{(positive)} \tag{6a}$$

$$\sum_{\mathbf{x}_\gamma} q_\gamma(\mathbf{x}_\gamma) = 1 \quad \forall_{\gamma \in \mathcal{R}} \qquad \textit{(normalized)} \tag{6b}$$

$$\sum_{\mathbf{x}_{\gamma'\setminus\gamma}} q_{\gamma'}(\mathbf{x}_{\gamma'}) = q_\gamma(\mathbf{x}_\gamma) \quad \forall_{\gamma,\gamma' \in \mathcal{R}; \gamma' \supset \gamma} \qquad \textit{(consistent)} \tag{6c}$$

Referring to the class of pseudo-marginals satisfying these constraints as $\mathcal{Q}$, we have the approximation

$$-\log Z \approx \min_{\mathbf{q} \in \mathcal{Q}} F_{\text{Kikuchi}}(\mathbf{q}) \, .$$

Furthermore, the hope is that the pseudo-marginals $q_\gamma(\mathbf{x}_\gamma)$ corresponding to this minimum are accurate approximations of the exact marginals $p_{\text{exact}}(\mathbf{x}_\gamma)$. The Kikuchi free energy and corresponding marginals are exact if the Hasse diagram turns out to be singly-connected (Pakzad & Anantharam, 2005).

2.5 Illustration

For illustration of the main concepts, we consider a probability model with 4 variables ("nodes") and pairwise interactions between each of the nodes as visualized in Figure 1(a). In obvious shorthand notation, the exact distribution is of the form

$$p_{\text{exact}}(\mathbf{x}) = \frac{1}{Z} \prod_{\{i,j\}} \psi_{ij}(x_i, x_j) = \frac{1}{Z} \psi_{12}\psi_{13}\psi_{14}\psi_{23}\psi_{24}\psi_{34} \, .$$

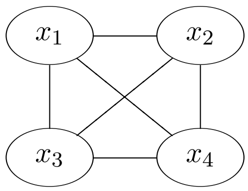
Note here that potentials originally defined on single nodes can always be incorporated in the definition of the two-node potentials. The region graph corresponding to the minimal choice of outer regions, i.e., equivalent to the potential subsets, is given in Figure 1(b). With the outer regions all pairs of nodes, the inner regions subsets are all single nodes. In fact, in this case the region graph is equivalent to a so-called factor graph (Kschischang, Frey, & Loeliger, 2001) and the Kikuchi approximation of the free energy boils down to a Bethe approximation:

$$\begin{aligned} F_{\text{Kikuchi}}(\mathbf{q}) \;\; = \;\; & -\sum_{\{i,j\}} \sum_{x_i,x_j} q_{ij}(x_i,x_j) \log \psi_{ij}(x_i,x_j) \\ & + \sum_{\{i,j\}} \sum_{x_i,x_j} q_{ij}(x_i,x_j) \log q_{ij}(x_i,x_j) + \sum_i (1-n_i) \sum_{x_i} q_i(x_i) \log q_i(x_i) \, , \end{aligned}$$

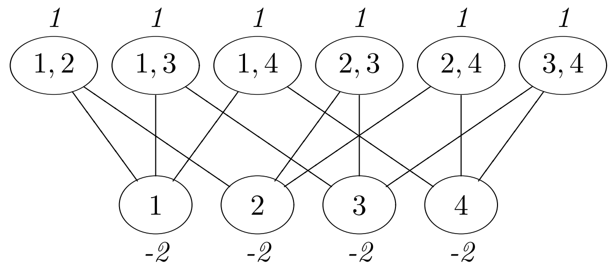
where $n_i = 3$ is the number of outer regions containing the inner region i.

The cluster variation method allows us to choose larger outer regions, for example, consisting of all triples $\{i,j,k\}$. We redefine the factorization of the potentials such that

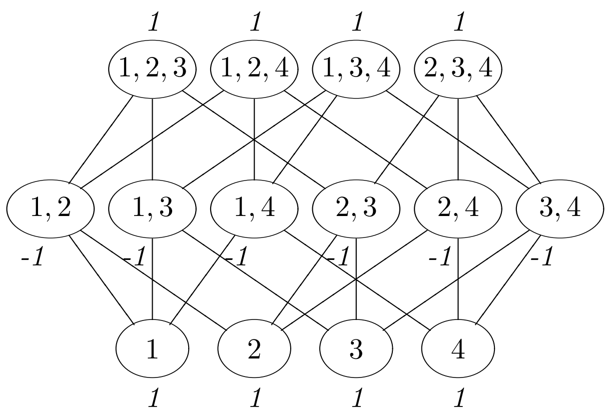
$$p_{\text{exact}}(\mathbf{x}) = \frac{1}{Z} \prod_{\{i,j,k\}} \psi_{ijk}(x_i,x_j,x_k) = \psi_{123}\psi_{124}\psi_{134}\psi_{234} \, ,$$



(a) Markov random field.



(b) Hasse diagram for the Bethe approximation.



(c) Region graph for the Kikuchi approximation.

Figure 1: Region graphs for the Bethe and Kikuchi approximations. Lines between nodes in the Markov random field (a) indicate edges. In the region graphs (b) and (c), the outer regions are drawn at the highest level. Lines indicate the "covering" relationship, where lower regions are covered by the higher regions. The oblique numbers are the overcounting numbers that follow from the Moebius formula. The Bethe approximation (b) corresponds to the minimal approximation with the outer regions equivalent to the cliques in the graph; here all pairs of nodes. The particular Kikuchi approximation (c) follows by taking for the outer regions all node triples.

for example through (distribute symmetrically)

$$\begin{aligned}\psi_{123} &\equiv [\psi_{12}\psi_{13}\psi_{23}]^{\frac{1}{2}} \\ \psi_{124} &\equiv [\psi_{12}\psi_{14}\psi_{24}]^{\frac{1}{2}} \\ \psi_{134} &\equiv [\psi_{13}\psi_{14}\psi_{34}]^{\frac{1}{2}} \\ \psi_{234} &\equiv [\psi_{23}\psi_{24}\psi_{34}]^{\frac{1}{2}} ,\end{aligned}$$

or through (assign to the first outer region)

$$\begin{aligned}\psi_{123} &\equiv \psi_{12}\psi_{13}\psi_{23} \\ \psi_{124} &\equiv \psi_{14}\psi_{24} \\ \psi_{134} &\equiv \psi_{34} \\ \psi_{234} &\equiv 1 .\end{aligned}$$

The corresponding region graph is given in Figure 1(c). Now the first-level inner regions are all pairs of nodes and the second-level inner regions are all single nodes, with overcounting numbers -1 and 1, respectively. The Kikuchi approximation of the entropy boils down to

$$S_{\text{Kikuchi}}(\mathbf{q}) = \sum_{\{i,j,k\}} S_{ijk} - \sum_{\{i,j\}} S_{ij} + \sum_{i} S_i .$$

The intuitive reasoning behind this approximation is as follows. The sum over all three-node entropies overcounts the two-node interactions (each combination $\{i,j\}$ appears twice rather than once), which therefore have to be discounted once. But now the single-node interactions are too much discounted (overcounting number -1 times 3 appearances, compared with the 3 appearances with overcounting number 1 in the three-node entropies), yielding the overcounting number $1 - 3 \times (1) - 3 \times (-1) = 1$.

2.6 Generalized and Loopy Belief Propagation

To summarize, finding the Kikuchi approximation of the partition function Z boils down to minimization of the Kikuchi free energy with respect to a set of pseudo-marginals under linear constraints between them. Introducing Lagrange multipliers for these constraints, it can be shown that fixed points of a popular algorithm called loopy belief propagation correspond to extrema of the Bethe free energy and, more generally, fixed points of generalized belief propagation to extrema of the Kikuchi free energy (Yedidia et al., 2001). However, these algorithms are not guaranteed to converge to a minimum and in practice do get stuck in for example limit cycles. This explains the search for convergent alternatives that directly minimize the Kikuchi free energy, which will be the topic of the rest of this paper.

3. Convexity of the Kikuchi Free Energy

In this section we will derive sufficient conditions for the Kikuchi free energy to be convex over the set of consistency constraints (6). This is relevant because if the Kikuchi free energy is indeed convex over the constraint set, it must have a unique minimum and the minimization problem is relatively straightforward. Furthermore, the argument that we will use in deriving these conditions will play an important role in the construction of efficient minimization algorithms later on.

3.1 Sufficient Conditions

We have to consider the Kikuchi free energy (5) as a function of the pseudo-marginals $\mathbf{q}$. In reasoning about convexity, we can disregard the energy term being linear in $\mathbf{q}$. The entropy terms give either a convex or a concave contribution, depending on whether the corresponding overcounting numbers are positive or negative, respectively. Ignoring the constraints (6), the free energy (5) is convex if and only if all concave contributions vanish, i.e., $c_\beta = 0$ for all $\beta \in \mathcal{R}_-$.

However, we really only care about the subspace induced by the constraints (6). Therefore we introduce the notion of *convexity over the set of constraints*. We call the free energy convex over the set of constraints (6) if

$$F(\lambda \mathbf{q}_1 + (1-\lambda)\mathbf{q}_2) \leq \lambda F(\mathbf{q}_1) + (1-\lambda)F(\mathbf{q}_2) \quad \forall_{0<\lambda<1} \forall_{\mathbf{q}_1, \mathbf{q}_2 \in \mathcal{Q}} \, .$$

Note that, since the constraints are all linear, if $\mathbf{q}_1$ and $\mathbf{q}_2$ satisfy the constraints (6), then so does $\lambda \mathbf{q}_1 + (1-\lambda)\mathbf{q}_2$. In the following, when we talk about convexity of the Kikuchi free energy, the conditioning on the constraint set is implicitly assumed.

One way to proceed is to make use of the (consistency) constraints to express the Kikuchi free energy in terms of the outer region pseudo-marginals only and then study its convexity. Our approach is along these lines. In particular, we will replace inner region pseudo-marginals that correspond to concave contributions by outer region pseudo-marginals. The pseudo-marginals corresponding to convex contributions are of no concern. In fact, we may be able to use these convex contributions as well to compensate for some of the concave contributions.

To make this reasoning more precise, we define *positive regions* (or perhaps better, nonnegative) $\gamma \in \mathcal{R}_+$, with $\mathcal{R}_+ \equiv \{\gamma \in \mathcal{R}; c_\gamma \geq 0\} \equiv \mathcal{O} \cup \mathcal{I}_+$ and *negative regions* $\beta \in \mathcal{R}_-$, with $\mathcal{R}_- \equiv \{\gamma \in \mathcal{R}; c_\gamma < 0\} \equiv \mathcal{I}_-$. The idea, formulated in the following theorem, is then that the Kikuchi free energy is convex if we can compensate the concave contributions of the negative regions $\mathcal{R}_-$ by the convex contributions of the positive regions $\mathcal{R}_+$.

Theorem 3.1. *The Kikuchi free energy is convex over the set of constraints (6) if there exists an "allocation matrix" $A_{\gamma\beta}$ between positive regions $\gamma \in \mathcal{R}_+$ and negative regions $\beta \in \mathcal{R}_-$ satisfying*

$$A_{\gamma\beta} \neq 0 \text{ only if } \gamma \supset \beta \qquad \text{(γ can be used to compensate β)} \tag{7a}$$

$$A_{\gamma\beta} \geq 0 \qquad \text{(positivity)} \tag{7b}$$

$$\sum_{\beta \subset \gamma} A_{\gamma\beta} \leq c_\gamma \quad \forall_{\gamma \in \mathcal{R}_+} \qquad \text{(sufficient amount of resources)} \tag{7c}$$

$$\sum_{\gamma \supset \beta} A_{\gamma\beta} \geq |c_\beta| \quad \forall_{\beta \in \mathcal{R}_-} \qquad \text{(sufficient compensation)} \tag{7d}$$

Proof First of all, we note that we do not have to worry about the energy terms that are linear in $\mathbf{q}$. In other words, to prove the theorem we can restrict ourselves to showing that minus the entropy

$$-S(\mathbf{q}) = -\left[\sum_{\gamma\in\mathcal{R}_+} c_\gamma S_\gamma(q_\gamma) - \sum_{\beta\in\mathcal{R}_-} |c_\beta| S_\beta(q_\beta)\right]$$

is convex over the set of constraints.

As an intermediate step, let us consider the combination of a convex entropy contribution of a positive region $\gamma \in \mathcal{R}_+$ with the concave entropy contribution of a negative inner region $\beta \in \mathcal{R}_-$, where β is a subset of γ:

$$\begin{aligned}
\Delta_{\gamma\beta}(\mathbf{q}) &\equiv -[S_\gamma(\mathbf{q}) - S_\beta(\mathbf{q})] = \sum_{\mathbf{x}_\gamma} q_\gamma(\mathbf{x}_\gamma)\log q_\gamma(\mathbf{x}_\gamma) - \sum_{\mathbf{x}_\beta} q_\beta(\mathbf{x}_\beta)\log q_\beta(\mathbf{x}_\beta) \\
&= \sum_{\mathbf{x}_\gamma} q_\gamma(\mathbf{x}_\gamma)\log q_\gamma(\mathbf{x}_\gamma) - \sum_{\mathbf{x}_\gamma} q_\gamma(\mathbf{x}_\gamma)\log q_\gamma(\mathbf{x}_\beta) \\
&= \sum_{\mathbf{x}_\beta} q_\gamma(\mathbf{x}_\beta)\left[\sum_{\mathbf{x}_{\gamma\backslash\beta}} q_\gamma(\mathbf{x}_{\gamma\backslash\beta}|\mathbf{x}_\beta)\log q_\gamma(\mathbf{x}_{\gamma\backslash\beta}|\mathbf{x}_\beta)\right],
\end{aligned}$$

where we used the standard definitions

$$q_\gamma(\mathbf{x}_\beta) \equiv \sum_{\mathbf{x}_{\gamma\backslash\beta}} q_\gamma(\mathbf{x}_\gamma) \quad\text{and}\quad q_\gamma(\mathbf{x}_{\gamma\backslash\beta}|\mathbf{x}_\beta) \equiv \frac{q_\gamma(\mathbf{x}_\gamma)}{q_\beta(\mathbf{x}_\beta)}\,.$$

In the first step, we applied the constraint $q_\beta(\mathbf{x}_\beta) = q_\gamma(\mathbf{x}_\beta)$ and extended the summation over $\mathbf{x}_\beta$ in the second term to a summation over $\mathbf{x}_\gamma$. In the second step we basically turned the difference between two entropies into (a weighted sum of) conditional entropies. The difference $\Delta_{\gamma\beta}$, which now only depends on q_γ, is, from Lemma A.1 in Appendix A, convex in q_γ. In other words, the concave contribution from S_β is fully compensated by the convex contribution S_γ, yielding an overall convex term in the relevant set of constraints.

The resulting operation is now a matter of resource allocation. For each concave contribution $|c_\beta|S_\beta$ we have to find convex contributions S_γ to compensate for it. Let $A_{\gamma\beta}$ denote the "amount of resources" that we take from positive region $\gamma \in \mathcal{R}_+$ to compensate for negative region $\beta \in \mathcal{R}_-$. Obviously, a positive region can only compensate negative regions that it contains, so $A_{\gamma\beta} = 0$ when β is not a subset of γ, which explains condition (7a). Now, in shorthand notation and with a little bit of rewriting

$$\begin{aligned}
-S(\mathbf{q}) &= -\left[\sum_{\gamma\in\mathcal{R}_+} c_\gamma S_\gamma - \sum_{\beta\in\mathcal{R}_-} |c_\beta| S_\beta\right] \\
&= -\sum_{\gamma\in\mathcal{R}_+}\left(c_\gamma - \sum_{\beta\subset\gamma} A_{\gamma\beta} + \sum_{\beta\subset\gamma} A_{\gamma\beta}\right) S_\gamma - \sum_{\beta\in\mathcal{R}_-}\left(-\sum_{\gamma\supset\beta} A_{\gamma\beta} + \sum_{\gamma\supset\beta} A_{\gamma\beta} - |c_\beta|\right) S_\beta \\
&= -\sum_{\gamma\in\mathcal{R}_+}\left(c_\gamma - \sum_{\beta\subset\gamma} A_{\gamma\beta}\right) S_\gamma - \sum_{\gamma\in\mathcal{R}_+}\sum_{\beta\subset\gamma} A_{\gamma\beta}[S_\gamma - S_\beta] - \sum_{\beta\in\mathcal{R}_-}\left[\sum_{\gamma\supset\beta} A_{\gamma\beta} - |c_\beta|\right] S_\beta\,.
\end{aligned}$$

Convexity of the first term is guaranteed if $c_\gamma - \sum_\beta A_{\gamma\beta} \geq 0$ (7c), of the second term if $A_{\gamma\beta} \geq 0$ (7b), and of the third term if $\sum_\gamma A_{\gamma\beta} - |c_\beta| \geq 0$ (7d). □

3.2 Checking the Conditions

Checking the conditions of Theorem 3.1 can be cast in the form of a linear programming problem, for example as follows. We define an auxiliary variable θ replacing condition (7c) by

$$\sum_{\gamma \supset \beta} A_{\gamma\beta} = \theta |c_\beta| \quad \forall_{\beta \in \mathcal{R}_-} \text{ \textit{(variable compensation)}} \tag{8}$$

Then we solve the linear programming problem that attempts to maximize the single variable θ under all constraints implied by the four conditions. The interpretation is that we try to use the available resources to compensate for as much of the concave contributions as we can. If we find a solution $\theta^* \geq 1$ all conditions are satisfied: the Kikuchi free energy is convex over the set of constraints and has a unique minimum. If the optimal θ^* turns out to be smaller than 1, there is no matrix A satisfying all constraints and convexity of the Kikuchi free energy is not guaranteed by Theorem 3.1.

Instead of solving the linear program, we can often get away with simpler checks. For example, we can guess a particular A and check whether the conditions (7) hold. An obvious choice is

$$A_{\gamma\beta} = \frac{c_\gamma}{n_\gamma^-} \text{ with } n_\gamma^- \equiv \sum_{\beta \in \mathcal{R}_-, \beta \subset \gamma} 1 \,,$$

which satisfies condition (7c) and when substituted into (7d) yields the condition

$$c_\beta + \sum_{\gamma \in \mathcal{R}_+, \gamma \supset \beta} \frac{c_\gamma}{n_\gamma^-} \geq 0 \quad \forall_{\beta \in \mathcal{R}_-} \,. \tag{9}$$

Similarly, the choice

$$A_{\gamma\beta} = \frac{|c_\beta|}{n_\beta^+} \text{ with } n_\beta^+ \equiv \sum_{\gamma \in \mathcal{R}_+, \gamma \supset \beta} 1$$

satisfies condition (7d) and yields the condition

$$\sum_{\beta \in \mathcal{R}_-, \beta \subset \gamma} \frac{c_\beta}{n_\beta^+} + c_\gamma \geq 0 \quad \forall_{\gamma \in \mathcal{R}_+} \tag{10}$$

when substituted into (7c). If (9) or (10) holds, Theorem 3.1 guarantees convexity of the Kikuchi free energy.

The above two conditions are sufficient, but not necessary for Theorem 3.1 to apply. A necessary condition is

$$\sum_{\beta \in \mathcal{R}_-} c_\beta + \sum_{\gamma \in \mathcal{R}_+} c_\gamma \geq 0 \tag{11}$$

which is easily derived by summing condition (7d) over all $\beta \in \mathcal{R}_-$ and substituting condition (7c). If condition (11) fails, we cannot use Theorem 3.1 to prove convexity of the Kikuchi free energy.

We would like to conjecture that the conditions in Theorem 3.1 are not only sufficient, but also necessary for convexity of the Kikuchi free energy. We will not pursue this any further here, because it is irrelevant for our current purposes. Furthermore, it may not be that relevant in practice either, since convexity by itself is a sufficient but not necessary condition for a unique minimum. Tatikonda and Jordan (2002), Heskes (2004), Ihler, Fisher, and Willsky (2005) give conditions for convergence of loopy belief propagation and uniqueness of the minimum of the corresponding Bethe free energy. These conditions do not only depend on the graphical structure, but also on the (strength of the) kernels $\psi_\alpha(\mathbf{x}_\alpha)$.

3.3 Related Work

Chiang and Forney (2001) present similar ideas, about convex entropy terms compensating concave terms in the set of constraints, and derive conditions for convexity of the Bethe free energy with pairwise potentials. The resulting conditions are formulated in terms of single-node marginals, which may be difficult both to validate in practice and to generalize to the Kikuchi case.

Closely related to our Theorem 3.1 is the following theorem of Pakzad and Anantharam (2002, 2005).

Theorem 3.2. (Pakzad & Anantharam, 2002, 2005) *The Kikuchi free energy (5) is convex over the set of consistency constraints imposed by a collection of regions R (and hence the constrained minimization problem has a unique solution) if the overcounting numbers c_γ and $c_{\gamma'}$ satisfy:*

$$\forall \mathcal{S} \subset \mathcal{R}, \quad \sum_{\gamma \in \mathcal{S}} c_\gamma + \sum_{\substack{\gamma' \in \mathcal{R} \setminus \mathcal{S}: \\ \exists \gamma \in \mathcal{S}, \gamma \subset \gamma'}} c_{\gamma'} \geq 0 \,. \tag{12}$$

In words, for any subset $\mathcal{S}$ of $\mathcal{R}$, the sum of overcounting numbers of elements of $\mathcal{S}$ and all their ancestors in $\mathcal{R}$ must be nonnegative.

In fact, using Hall's (1935) matching theorem, it can be shown that the conditions (7) in our Theorem 3.1 are equivalent to the conditions (12) in Theorem 3.2. The latter are more direct and do not require the solution of a linear program.

Both Theorem 3.1 and Theorem 3.2 can be used to show that the Bethe free energy for graphs with a single loop is convex over the set of constraints (Heskes, 2004; McEliece & Yildirim, 2003; Pakzad & Anantharam, 2002, 2005).

3.4 Minimization of the Convex Kikuchi Free Energy

If the Kikuchi free energy is convex, it is not only guaranteed to have a unique minimum, but this minimum is also relatively easy to find with a message-passing algorithm similar to standard (loopy) belief propagation.

The basic idea is as follows. We here focus on the case in which all overcounting numbers are positive. The case with negative overcounting numbers is more involved and worked out in Appendix B. Furthermore, here and in the rest of this paper we ignore the positivity constraints (6a). It is easy to check that these are satisfied at the solutions we obtain. We introduce Lagrange multipliers $\lambda_{\gamma'\gamma}(x_\gamma)$ for the consistency constraints as well as λ_γ for the normalization constraints and construct the Lagrangian

$$L(\mathbf{q},\boldsymbol{\lambda}) = F_{\text{Kikuchi}}(\mathbf{q}) + \sum_{\substack{\gamma',\gamma \\ \gamma\subset\gamma'}} \sum_{\mathbf{x}_\gamma} \lambda_{\gamma'\gamma}(\mathbf{x}_\gamma)\left[q_\gamma(\mathbf{x}_\gamma) - \sum_{\mathbf{x}_{\gamma'\setminus\gamma}} q_{\gamma'}(\mathbf{x}_{\gamma'})\right] + \sum_\gamma \lambda_\gamma \left[1 - \sum_{\mathbf{x}_\gamma} q_\gamma(\mathbf{x}_\gamma)\right] . \tag{13}$$

Minimization of the Kikuchi free energy under the appropriate consistency and normalization constraints is, in terms of this Lagrangian, equivalent to

$$\min_{\mathbf{q}\in\mathcal{Q}} F_{\text{Kikuchi}}(\mathbf{q}) = \min_{\mathbf{q}} \max_{\boldsymbol{\lambda}} L(\mathbf{q},\boldsymbol{\lambda}) ,$$

where the minimization over $\mathbf{q}$ is now unconstrained. Standard results from constrained optimization (e.g., Luenberger, 1984) tell us that

$$\min_{\mathbf{q}} \max_{\boldsymbol{\lambda}} L(\mathbf{q},\boldsymbol{\lambda}) \geq \max_{\boldsymbol{\lambda}} \min_{\mathbf{q}} L(\mathbf{q},\boldsymbol{\lambda}) ,$$

with equality for convex problems under linear equality constraints. That is, for convex problems we are allowed to interchange the maximum over $\boldsymbol{\lambda}$ and the minimum over $\mathbf{q}$. Furthermore, the optimal $\mathbf{q}^*(\boldsymbol{\lambda})$ corresponding to the minimum of the Lagrangian (13) as a function of $\boldsymbol{\lambda}$ is unique, since $L(\mathbf{q},\boldsymbol{\lambda})$ is convex in $\mathbf{q}$ for all $\boldsymbol{\lambda}$. Substitution of the solution then yields the so-called dual

$$L^*(\boldsymbol{\lambda}) \equiv \min_{\mathbf{q}} L(\mathbf{q},\boldsymbol{\lambda}) = L(\mathbf{q}^*(\boldsymbol{\lambda}),\boldsymbol{\lambda}) . \tag{14}$$

This dual is concave in $\boldsymbol{\lambda}$ and has a unique maximum.

Many algorithms can be used to find the maximum of the dual (14). A particular one, derived in Appendix B, is given in Algorithm 1. It slightly differs from those presented by Yedidia et al. (2005) and Yuille (2002) by sending messages (messages are directly related to Lagrange multipliers) only between inner regions and outer regions, i.e., never between inner regions subsets and other inner regions. The price one has to pay is that the update in line 7 depends on the overcounting number c_β. For the Bethe free energy, with $c_\beta = 1 - n_\beta$, we obtain the standard (loopy) belief propagation update rules. The particular ordering in Algorithm 1, running over inner regions and updating the messages between an inner region and all its neighboring outer regions, guarantees that the dual (14) increases at each iteration[1]. The local partition functions Z_α and Z_β in lines 10 and 7 are chosen such as to normalize the pseudo-marginals $q_\alpha(\mathbf{x}_\alpha)$ and $q_\beta(\mathbf{x}_\beta)$. This normalization is not strictly necessary, but helps to prevent numerical instability. Algorithm 1 can be initialized by setting all messages $\mu_{\alpha\to\beta}(\mathbf{x}_\beta) = 1$ and skipping lines 3 to 6 at the first iteration.

1. For positive overcounting numbers c_β. The argumentation with negative overcounting numbers is more complicated and may require damping of the updates to achieve convergence. See Appendix B for details.

Algorithm 1 Message-passing algorithm for constrained minimization of a Kikuchi free energy.

while $\neg$converged **do**
  for all $\beta \in \mathcal{I}$ **do**
    for all $\alpha \in \mathcal{O}, \alpha \supset \beta$ **do**
      $$q_\alpha(\mathbf{x}_\beta) = \sum_{\mathbf{x}_{\alpha \setminus \beta}} q_\alpha(\mathbf{x}_\alpha)$$
      $$\mu_{\alpha \to \beta}(\mathbf{x}_\beta) = \frac{q_\alpha(\mathbf{x}_\beta)}{\mu_{\beta \to \alpha}(\mathbf{x}_\beta)}$$
    end for
    $$q_\beta(\mathbf{x}_\beta) = \frac{1}{Z_\beta} \prod_{\substack{\alpha \in \mathcal{O}, \\ \alpha \supset \beta}} \mu_{\alpha \to \beta}(\mathbf{x}_\beta)^{\frac{1}{n_\beta + c_\beta}}$$
    for all $\alpha \in \mathcal{O}, \alpha \supset \beta$ **do**
      $$\mu_{\beta \to \alpha}(\mathbf{x}_\beta) = \frac{q_\beta(\mathbf{x}_\beta)}{\mu_{\alpha \to \beta}(\mathbf{x}_\beta)}$$
      $$q_\alpha(\mathbf{x}_\alpha) = \frac{1}{Z_\alpha} \psi_\alpha(\mathbf{x}_\alpha) \prod_{\substack{\beta \in \mathcal{I}, \\ \beta \subset \alpha}} \mu_{\beta \to \alpha}(\mathbf{x}_\beta)$$
    end for
  end for
end while

4. Double-Loop Algorithms for Guaranteed Convergence

Even when the Kikuchi free energy is not convex, we can still run Algorithm 1 in the hope that it converges to a fixed point. This fixed point then must correspond to an extremum of the Kikuchi free energy under the appropriate constraints (Yedidia et al., 2001). Even better, empirically for the general Kikuchi free energy and provably for the Bethe free energy (Heskes, 2003), this extremum is in fact a minimum. However, in practice this single-loop[2] algorithm does not always converge and we have to resort to double-loop algorithms to guarantee convergence to a minimum of the Kikuchi free energy.

4.1 The General Procedure

We introduce a class of such double-loop algorithms based on the following theorem.

2. Note that "single loop" here refers to the message-passing algorithm and has nothing to do with the notion of a single loop in the graphical model.

Theorem 4.1. *Given a function $F_{\text{convex}}(\mathbf{q};\mathbf{q}')$ with properties*

$$F_{\text{convex}}(\mathbf{q};\mathbf{q}') \geq F_{\text{Kikuchi}}(\mathbf{q}) \qquad \forall_{\mathbf{q},\mathbf{q}' \in \mathcal{Q}} \qquad \textit{(bound)} \tag{15a}$$

$$F_{\text{convex}}(\mathbf{q};\mathbf{q}) = F_{\text{Kikuchi}}(\mathbf{q}) \text{ and} \quad \left.\frac{\partial F_{\text{convex}}(\mathbf{q};\mathbf{q}')}{\partial \mathbf{q}}\right|_{\mathbf{q}'=\mathbf{q}} = \frac{\partial F_{\text{Kikuchi}}(\mathbf{q})}{\partial \mathbf{q}} \qquad \forall_{\mathbf{q} \in \mathcal{Q}} \qquad \textit{(touching)} \tag{15b}$$

$$F_{\text{convex}}(\mathbf{q};\mathbf{q}') \textit{ is convex in } \mathbf{q} \in \mathcal{Q} \qquad \forall_{\mathbf{q}' \in \mathcal{Q}} \qquad \textit{(convex)} \tag{15c}$$

the algorithm

$$\mathbf{q}_{n+1} = \operatorname*{argmin}_{\mathbf{q} \in \mathcal{Q}} F_{\text{convex}}(\mathbf{q};\mathbf{q}_n) \,, \tag{16}$$

with $\mathbf{q}_n$ the pseudo-marginals at iteration n, is guaranteed to converge to a local minimum of the Kikuchi free energy $F_{\text{Kikuchi}}(\mathbf{q})$ under the appropriate constraints.

Proof It is immediate that the Kikuchi free energy decreases with each iteration:

$$F_{\text{Kikuchi}}(\mathbf{q}_{n+1}) \leq F_{\text{convex}}(\mathbf{q}_{n+1};\mathbf{q}_n) \leq F_{\text{convex}}(\mathbf{q}_n;\mathbf{q}_n) = F_{\text{Kikuchi}}(\mathbf{q}_n) \,,$$

where the first inequality follows from condition (15a) (upper bound) and the second from the definition of the algorithm. The gradient property (15b) ensures that the algorithm is only stationary in points where the gradient of F_{Kikuchi} is zero. By construction $\mathbf{q}_n \in \mathcal{Q}$ for all n. $\square$

See Figure 2 for an illustration of the algorithm and the proof. In fact, the convexity of F_{convex} has not been used to establish the proof. But, as argued in Section 3.4, from an algorithmic point of view constrained minimization of a convex functional is much simpler than constrained minimization of a non-convex functional. This general idea, replacing the minimization of a complex functional by the consecutive minimization of an easier to handle upper bound of this functional, forms the basis of popular algorithms such as the EM algorithm (Dempster, Laird, & Rubin, 1977; Neal & Hinton, 1998) and iterative scaling/iterative proportional fitting (Darroch & Ratcliff, 1972; Jiroušek & Přeučil, 1995). Intuitively, the tighter the bound, the faster the algorithm.

4.2 Bounding the Concave Terms

As a first step, to lay out the main ideas, we build a convex bound by removing all concave entropy contributions for $\beta \in \mathcal{I}_-$. To do so, we will make use of the linear bound

$$-\sum_{\mathbf{x}_\beta} q_\beta(\mathbf{x}_\beta) \log q_\beta(\mathbf{x}_\beta) \leq -\sum_{\mathbf{x}_\beta} q_\beta(\mathbf{x}_\beta) \log q'_\beta(\mathbf{x}_\beta) \,, \tag{17}$$

which directly follows from

$$0 \leq \text{KL}(q_\beta, q'_\beta) = \sum_{\mathbf{x}_\beta} q_\beta(\mathbf{x}_\beta) \log \left[\frac{q_\beta(\mathbf{x}_\beta)}{q'_\beta(\mathbf{x}_\beta)}\right]$$

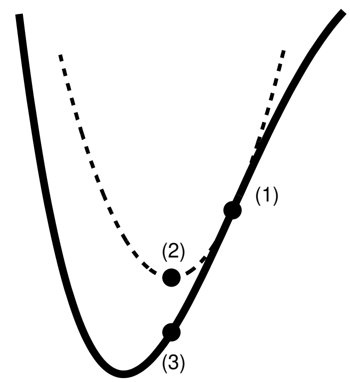


Figure 2: Illustration of the proposed algorithm and corresponding convergence proof. At iteration n, $F_{\text{convex}}(\mathbf{q};\mathbf{q}_n)$ (dashed line) is a convex bound of the non-convex $F_{\text{Kikuchi}}(\mathbf{q})$ (solid line). They touch at $\mathbf{q}_n$, point (1), where $F_{\text{convex}}(\mathbf{q}_n;\mathbf{q}_n) = F_{\text{Kikuchi}}(\mathbf{q}_n)$. At the minimum, point (2), we have $F_{\text{convex}}(\mathbf{q}_{n+1};\mathbf{q}_n) \leq F_{\text{convex}}(\mathbf{q}_n;\mathbf{q}_n)$. The corresponding Kikuchi free energy, point (3), obeys $F_{\text{Kikuchi}}(\mathbf{q}_{n+1}) \leq F_{\text{convex}}(\mathbf{q}_{n+1};\mathbf{q}_n)$ because of the bounding property.

with KL the Kullback-Leibler divergence. Our choice F_{convex} then reads

$$
\begin{aligned}
F^{(1)}_{\text{convex}}(\mathbf{q};\mathbf{q}') \;=\;& \sum_\alpha \sum_{\mathbf{x}_\alpha} q_\alpha(\mathbf{x}_\alpha) \log\left[\frac{q_\alpha(\mathbf{x}_\alpha)}{\psi_\alpha(\mathbf{x}_\alpha)}\right] + \sum_{\beta\in\mathcal{I}_+} c_\beta \sum_{\mathbf{x}_\beta} q_\beta(\mathbf{x}_\beta)\log q_\beta(\mathbf{x}_\beta) \\
& - \sum_{\beta\in\mathcal{I}_-} |c_\beta| \sum_{\mathbf{x}_\beta} q_\beta(\mathbf{x}_\beta) \log q'_\beta(\mathbf{x}_\beta) + \sum_{\beta\in\mathcal{I}_-} |c_\beta| \left[1 - \sum_{\mathbf{x}_\beta} q_\beta(\mathbf{x}_\beta)\right] . \quad (18)
\end{aligned}
$$

It is easy to check that this functional has properties (15a) and (15c). The last term has been added to fulfill property (15b). Next we make the crucial observation that, using the constraints (6) and for fixed $\mathbf{q}'$, we can rewrite $F^{(1)}_{\text{convex}}$ in the "normal form" (5):

$$
F^{(1)}_{\text{convex}}(\mathbf{q};\mathbf{q}') = \sum_\alpha \sum_{\mathbf{x}_\alpha} q_\alpha(\mathbf{x}_\alpha) \log\left[\frac{q_\alpha(\mathbf{x}_\alpha)}{\tilde{\psi}_\alpha(\mathbf{x}_\alpha)}\right] + \sum_\beta \tilde{c}_\beta \sum_{\mathbf{x}_\beta} q_\beta(\mathbf{x}_\beta) \log q_\beta(\mathbf{x}_\beta) + C(\mathbf{q}) \,, \quad (19)
$$

where $C(\mathbf{q})$ evaluates to zero for all $\mathbf{q} \in \mathcal{Q}$ and where $\tilde{\psi}$, which implicitly depends on $\mathbf{q}'$, and $\tilde{c}$ are defined through

$$
\log \tilde{\psi}_\alpha(\mathbf{x}_\alpha) \equiv \log \psi_\alpha(\mathbf{x}_\alpha) + \sum_{\substack{\beta\in\mathcal{I}_-,\\ \beta\subset\alpha}} \frac{|c_\beta|}{n_\beta} \log q'_\beta(\mathbf{x}_\beta) \;\text{ and }\; \tilde{c}_\beta \equiv \begin{cases} 0 & \forall_{\beta\in\mathcal{I}_-} \\ c_\beta & \forall_{\beta\in\mathcal{I}_+} \end{cases} . \quad (20)
$$

That is, we can always to incorporate the terms now linear in q_β in the energy term by redefinition of the potentials. Here we have chosen to distribute each of these terms equally over the n_β neighboring outer regions, but other choices are possible as well.

The term $C(\mathbf{q})$ in (19) evaluates to zero for all $\mathbf{q} \in \mathcal{Q}$ and is thus irrelevant to the optimization in the inner loop. It consists of terms such as the last one in (18) that only serve to make the bound $F^{(1)}_{\text{convex}}$ satisfy (15b). In the construction of other bounds below, we will ignore such terms: they do not affect the algorithm in any way[3].

Now that we have $F^{(1)}_{\text{convex}}$ both convex and in normal form, we can use Algorithm 1 to solve the constrained problem (16). The resulting double-loop algorithm can be described in two lines.

Outer loop: recompute $\tilde{\psi}$ from (20) with $\mathbf{q}' = \mathbf{q}_n$.

Inner loop: run Algorithm 1 with $\tilde{\psi}$ for ψ and $\tilde{c}$ for c, yielding $\mathbf{q}_{n+1}$.

In each inner loop, we can initialize the messages to the converged values of the previous inner loop.

4.3 Bounding the Convex Terms

In this section we will show that in many cases we can make the algorithm both better and simpler. The idea is to bound not only the concave, but also the convex entropy contributions from inner regions. That is, we enforce $\tilde{c}_\beta \equiv 0 \;\; \forall \beta \in \mathcal{I}$ and set

$$F^{(2)}_{\text{convex}}(\mathbf{q}; \mathbf{q}') = \sum_\alpha \sum_{\mathbf{x}_\alpha} q_\alpha(\mathbf{x}_\alpha) \log \left[\frac{q_\alpha(\mathbf{x}_\alpha)}{\tilde{\psi}_\alpha(\mathbf{x}_\alpha)} \right] , \tag{21}$$

with now

$$\log \tilde{\psi}_\alpha(\mathbf{x}_\alpha) \equiv \log \psi_\alpha(\mathbf{x}_\alpha) - \sum_{\beta \subset \alpha} \frac{c_\beta}{n_\beta} \log q'_\beta(\mathbf{x}_\beta) . \tag{22}$$

Let us first explain why the algorithm based on $F^{(2)}_{\text{convex}}$ is simpler than the one based on $F^{(1)}_{\text{convex}}$. In (21), all reference to inner regions has disappeared. In fact, the only constraints that we have to care about are that the outer regions pseudo-marginals should agree on their intersections. Consequently, in the inner loop (Algorithm 1), we only have to run over those inner regions β that are direct intersections of the outer regions, that is, those β for which there exist outer regions α and α' such that $\mathbf{x}_\beta = \mathbf{x}_\alpha \cap \mathbf{x}_{\alpha'}$. Similar arguments can be used for the algorithm based on (19) as well, neglecting all negative inner regions $\beta \in \mathcal{I}_-$ that do not correspond to direct intersections of outer regions. In practice, however, most negative inner regions *are* direct intersections of the outer regions, whereas many positive inner regions arise at the next level, from intersections of intersections. See for instance the example of Figure 1, where all six negative inner regions are direct intersections of outer regions, in contrast with all four positive inner regions.

From (17), but now applied to the positive inner regions, it is clear that $F^{(2)}_{\text{convex}}(\mathbf{q}; \mathbf{q}') \leq F^{(1)}_{\text{convex}}(\mathbf{q}; \mathbf{q}')$: when it is a bound, $F^{(2)}_{\text{convex}}$ is a tighter bound than $F^{(1)}_{\text{convex}}$ and we can expect the algorithm based on $F^{(2)}_{\text{convex}}$ to perform better. It remains to be shown under which conditions $F_{\text{Kikuchi}}(\mathbf{q}) \leq F^{(2)}_{\text{convex}}(\mathbf{q}; \mathbf{q}')$. This is where the following theorem comes in.

3. Alternatively, we could relax condition (15b) to the statement that the gradients of F_{convex} and F_{Kikuchi} only have to be equal in the subspace orthogonal to the constraints. With this milder condition, $C(\mathbf{q})$ as well as the last term (18) are no longer needed.

Theorem 4.2. *The functional F_{convex} in (21) is a convex bound of the Kikuchi free energy (5) if there exists an "allocation matrix" $A_{\gamma\beta}$ between negative inner regions $\gamma \in \mathcal{I}_-$ and positive inner regions $\beta \in \mathcal{I}_+$ satisfying*

$$A_{\gamma\beta} \neq 0 \text{ only if } \gamma \supset \beta \qquad (\gamma \text{ can be used to compensate } \beta) \tag{23a}$$

$$A_{\gamma\beta} \geq 0 \qquad (\text{positivity}) \tag{23b}$$

$$\sum_{\beta \subset \gamma} A_{\gamma\beta} \leq |c_\gamma| \quad \forall_{\gamma \in \mathcal{I}_-} \qquad (\text{sufficient amount of resources}) \tag{23c}$$

$$\sum_{\gamma \supset \beta} A_{\gamma\beta} \geq c_\beta \quad \forall_{\beta \in \mathcal{I}_+} \qquad (\text{sufficient compensation}) \tag{23d}$$

Proof Not surprisingly, the proof follows the same line of reasoning as the proof of Theorem 3.1. First we consider the combination of a concave entropy contribution from $\gamma \in \mathcal{I}_-$ as in (17) with a convex entropy contribution from $\beta \in \mathcal{I}_+, \beta \subset \gamma$:

$$\begin{aligned} -\sum_{\mathbf{x}_\gamma} q_\gamma(\mathbf{x}_\gamma) \log q_\gamma(\mathbf{x}_\gamma) + \sum_{\mathbf{x}_\beta} q_\beta(\mathbf{x}_\beta) \log q_\beta(\mathbf{x}_\beta) \leq \\ -\sum_{\mathbf{x}_\gamma} q_\gamma(\mathbf{x}_\gamma) \log q'_\gamma(\mathbf{x}_\gamma) + \sum_{\mathbf{x}_\beta} q'_\beta(\mathbf{x}_\beta) \log q'_\beta(\mathbf{x}_\beta) \,, \end{aligned} \tag{24}$$

which follows from

$$\begin{aligned} 0 \quad &\leq \quad \sum_{\mathbf{x}_\gamma} q_\beta(\mathbf{x}_\beta) \left\{ q_\gamma(\mathbf{x}_{\gamma\backslash\beta}|\mathbf{x}_\beta) \left[\frac{q_\gamma(x_{\gamma\backslash\beta}|\mathbf{x}_\beta)}{q'_\gamma(\mathbf{x}_{\gamma\backslash\beta}|\mathbf{x}_\beta)} \right] \right\} \\ &= \quad \sum_{\mathbf{x}_\gamma} q_\gamma(\mathbf{x}_\beta) \left\{ \frac{q_\gamma(\mathbf{x}_\gamma)}{q_\gamma(\mathbf{x}_\beta)} \log \left[\frac{q_\gamma(\mathbf{x}_\gamma)}{q_\gamma(\mathbf{x}_\beta)} \frac{q'_\gamma(\mathbf{x}_\beta)}{q'_\gamma(\mathbf{x}_\gamma)} \right] \right\} \,, \end{aligned}$$

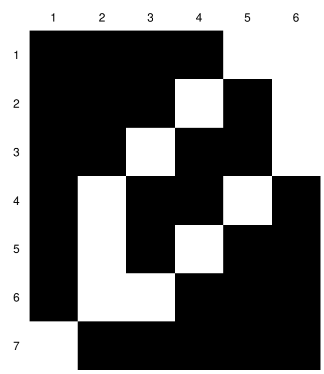
where we recognize the term between braces as a Kullback-Leibler divergence between two probability distributions.

To show that the difference between $F^{(2)}_{\text{convex}}$ and F_{Kikuchi} is nonnegative, we should be able to compensate each of the concave contributions c_β for all $\beta \in \mathcal{I}_+$ with convex contributions from $\gamma \in \mathcal{I}_-$ with $\gamma \supset \beta$, without exceeding the available amount of "resources" $|c_\gamma|$. In shorthand notation, with

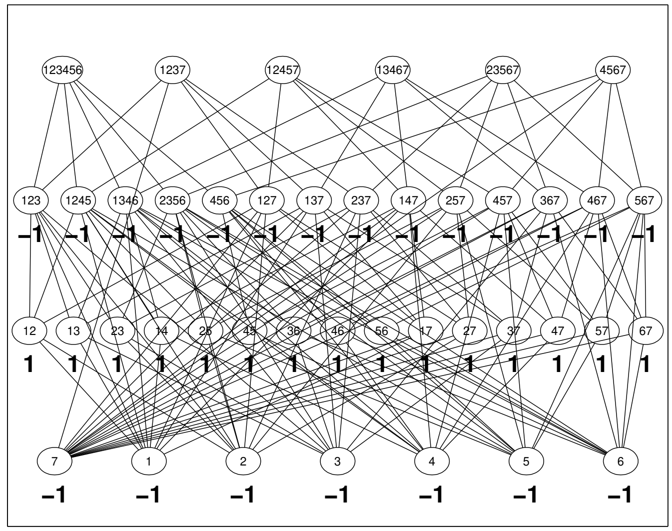
$$K_\beta \equiv \sum_{\mathbf{x}_\beta} q_\beta(\mathbf{x}_\beta) \log \left[\frac{q_\beta(\mathbf{x}_\beta)}{q'_\beta(\mathbf{x}_\beta)} \right] \,,$$

we have the decomposition

$$\begin{aligned} F^{(2)}_{\text{convex}} - F_{\text{Kikuchi}} &= -\sum_{\beta \in \mathcal{I}} c_\beta K_\beta = \sum_{\gamma \in \mathcal{I}_-} |c_\gamma| K_\gamma - \sum_{\beta \in \mathcal{I}_+} c_\beta K_\beta \\ &= \sum_{\gamma \in \mathcal{I}_-} \left(|c_\gamma| - \sum_{\beta \subset \gamma} A_{\gamma\beta} \right) K_\gamma + \sum_{\gamma \in \mathcal{I}_-} \sum_{\beta \subset \gamma} A_{\gamma\beta}(K_\gamma - K_\beta) + \sum_{\beta \in \mathcal{I}_+} \left(\sum_{\gamma \supset \beta} A_{\gamma\beta} - c_\beta \right) K_\beta \geq 0 \,, \end{aligned}$$



(a) Outer regions.



(b) Region graph.

Figure 3: Smallest example showing that the conditions for Theorem 4.2 need not always hold for a region graph and overcounting numbers constructed with the cluster variation method. (a) Visualization of the outer regions: black means that the variable (1 to 7) is part of the outer region (1 to 6).

(b) Region graph with overcounting numbers in boldface. The positive overcounting numbers at the third level just outweigh the negative overcounting numbers at the second level.

where the inequality follows since each of the terms itself is guaranteed to be nonnegative when the conditions (23) are satisfied. $\square$

As above, the conditions of Theorem 4.2 can be checked with a linear program. Having generated many different sets of overcounting numbers resulting from the Moebius formula (4), we started wondering whether the conditions (23) are perhaps automatically satisfied. However, exhaustively checking all possible outer region combinations given a fixed number of variables, we did come up with a counterexample. The smallest counterexample that violates the conditions for Theorem 4.2, is illustrated in Figure 3.

Even if, as in this counterexample, not all positive inner regions can be compensated for by negative inner regions, it will pay to get rid of as many as possible. Finding the optimal assignment may be a complex problem, but heuristics are easy to find (see Appendix C).

4.4 Pulling Out a Tree or More

In the previous section we tightened the convex bound $F^{(1)}_{\text{convex}}$ of the Kikuchi free energy F_{Kikuchi} by bounding convex contributions from positive regions as well. Another way to get a tighter bound is to bound only part of the concave contributions from the negative inner regions. We will first illustrate this by considering the Bethe free energy, i.e., just non-overlapping negative inner regions (nodes) with $c_\beta = 1 - n_\beta$.

The Bethe free energy is convex for singly-connected structures. Inspired by Teh and Welling (2002), we choose a set of nodes $\beta \in \mathcal{I}_{\text{bound}}$ such that the remaining nodes $\beta \in \mathcal{I}_{\text{free}}$ become singly-connected and take

$$\begin{aligned} F^{(3)}_{\text{convex}}(\mathbf{q};\mathbf{q}') &= \sum_\alpha \sum_{\mathbf{x}_\alpha} q_\alpha(\mathbf{x}_\alpha) \log\left[\frac{q_\alpha(\mathbf{x}_\alpha)}{\psi_\alpha(\mathbf{x}_\alpha)}\right] + \sum_{\beta \in \mathcal{I}_{\text{free}}} (1-n_\beta) \sum_{\mathbf{x}_\beta} q_\beta(\mathbf{x}_\beta) \log q_\beta(\mathbf{x}_\beta) \\ &\quad + \sum_{\beta \in \mathcal{I}_{\text{bound}}} (1-n_\beta) \sum_{\mathbf{x}_\beta} q_\beta(\mathbf{x}_\beta) \log q'_\beta(\mathbf{x}_\beta) . \end{aligned} \tag{25}$$

That is, we bound the entropy terms corresponding to the "bounded" nodes $\beta \in \mathcal{I}_{\text{bound}}$ and simply keep the entropy terms correspond to the "free" nodes $\beta \in \mathcal{I}_{\text{free}}$. By construction F_{convex} satisfies all conditions (15). Furthermore, it can be rewritten in the normal form (5) with definitions

$$\log \tilde{\psi}_\alpha(\mathbf{x}_\alpha) \equiv \log \psi_\alpha(\mathbf{x}_\alpha) - \sum_{\substack{\beta \in \mathcal{I}_{\text{bound}}, \\ \beta \subset \alpha}} \frac{1-n_\beta}{n_\beta} \log q'_\beta(\mathbf{x}_\beta) \quad \text{and} \quad \tilde{c}_\beta \equiv \begin{cases} 0 & \forall_{\beta \in \mathcal{I}_{\text{bound}}} \\ 1-n_\beta & \forall_{\beta \in \mathcal{I}_{\text{free}}} \end{cases} .$$

Note that the resulting inner-loop algorithm is *not* completely equivalent to running standard belief propagation on the tree of all free nodes: we do have to send messages to and from the bounded nodes $\beta \in \mathcal{I}_{\text{bound}}$ as well to enforce the constraints $q_\alpha(\mathbf{x}_\beta) = q_{\alpha'}(\mathbf{x}_\beta)$ for $\alpha, \alpha' \supset \beta$.

Rather than pulling out a single tree, we can also pull out "a convex combination of trees". That is, suppose that we have several bounds, each of them the result of pulling out a particular tree and with a corresponding set of overcounting numbers $\tilde{c}^i$. Then any convex combination

$$\tilde{c}_\beta = \sum_i w_i \tilde{c}^i_\beta \quad \text{with} \quad w_i \geq 0 \quad \text{and} \quad \sum_i w_i = 1$$

also corresponds to a convex bound. More generally, we can combine the ideas in this and the previous section by choosing $\tilde{c}_\beta$ such that the resulting bound is just convex. A procedure for doing so is given in Appendix C. Basically, we first try to shield as much of the concave entropy contributions by convex entropy contributions as we can. Next, we tighten the bound further by incorporating convex contributions in the linear bounds of the concave contributions that we did not manage to shield in the first step. Both steps can be cast in the form of an easy to solve linear programming problem.

4.5 Related Work

The double-loop algorithm described in Section 4.2 and based on $F^{(1)}_{\text{convex}}$ is closely related to Yuille's (2002) CCCP (concave-convex procedure) algorithm. Although originally formulated in a completely different way, CCCP applied for minimization of the Kikuchi free energy can also be understood as a particular case of the general procedure outlined in Theorem 4.1. More specifically, it is based on bounding the concave contributions with

$$-|c_\beta| \sum_{\mathbf{x}_\beta} q_\beta(\mathbf{x}_\beta) \log q_\beta(\mathbf{x}_\beta) \leq \sum_{\mathbf{x}_\beta} q_\beta(\mathbf{x}_\beta) \log q_\beta(\mathbf{x}_\beta) - (|c_\beta| - 1) \sum_{\mathbf{x}_\beta} q_\beta(\mathbf{x}_\beta) \log q'_\beta(\mathbf{x}_\beta) , \tag{26}$$

which is to be compared with (17). That is, before bounding the concave entropy contributions, part of these concave terms are taken over to the "convex side". The reason for doing so is that the CCCP algorithm requires the functional to be convex, independent of the constraints involved[4]. Our procedure, on the other hand, makes use of the fact that the functional only has to be convex over the set of constraints. This allows us to use tighter bounds, yielding more efficient and sometimes simpler algorithms. On a less important note, the inner-loop algorithm and in particular the message-passing scheme applied by Yuille (2002) is somewhat different.

The double-loop algorithm based on $F^{(3)}_{\text{convex}}$ in (25) is inspired by Teh and Welling's (2002) UPS (unified propagation and scaling) algorithm. The difference is that where we bound the entropy contributions from nodes on the tree, in UPS these nodes (and thus the entropy contributions) are *clamped* to the values resulting from the previous inner loop. That is, each inner loop in the UPS algorithm corresponds to minimizing

$$
\begin{aligned}
F^{\text{UPS}}_{\text{convex}}(\mathbf{q};\mathbf{q}') \;=\; & \sum_\alpha \sum_{\mathbf{x}_\alpha} q_\alpha(\mathbf{x}_\alpha) \log\left[\frac{q_\alpha(\mathbf{x}_\alpha)}{\psi_\alpha(\mathbf{x}_\alpha)}\right] + \sum_{\beta\in\mathcal{I}_{\text{free}}} (1-n_\beta) \sum_{\mathbf{x}_\beta} q_\beta(\mathbf{x}_\beta)\log q_\beta(\mathbf{x}_\beta) \\
& - \sum_{\beta\in\mathcal{I}_{\text{clamped}}} (1-n_\beta) \sum_{\mathbf{x}_\beta} q'_\beta(\mathbf{x}_\beta)\log q'_\beta(\mathbf{x}_\beta)\,.
\end{aligned}
$$

under the constraints

$$
q_\alpha(\mathbf{x}_\beta) = q_\beta(\mathbf{x}_\beta) \;\; \forall_{\beta\in\mathcal{I}_{\text{free}},\alpha\supset\beta}\;,\;\; \text{yet} \;\; q_\alpha(\mathbf{x}_\beta) = q'_\beta(\mathbf{x}_\beta) \;\; \forall_{\beta\in\mathcal{I}_{\text{clamped}},\alpha\supset\beta}\,.
$$

This boils down to an iterative scaling algorithm, which is also relatively easy to solve. At each outer-loop iteration, a different choice is made for $\mathcal{I}_{\text{free}}$ and $\mathcal{I}_{\text{clamped}}$. The UPS algorithm can be understood as coordinate descent and is guaranteed to converge to a local minimum of Bethe free energy (under appropriate conditions on the choices made for $\mathcal{I}_{\text{free}}$ and $\mathcal{I}_{\text{clamped}}$). The inner loop that results from $F^{(3)}_{\text{convex}}$ also allows for changes in the marginals $q_\beta(\mathbf{x}_\beta)$ for $\beta \in \mathcal{I}_{\text{bound}}$, i.e., is more flexible and can make larger steps. Loosely speaking, $F^{(3)}_{\text{convex}}$ is again a tighter bound than $F^{\text{UPS}}_{\text{convex}}$. Furthermore, in our approach we can but do not have to choose different subdivisions between "bounded" and "free" nodes within each inner loop.

Wainwright, Jaakkola, and Willsky (2002b, 2002a) present similar ideas, exploiting the convexity of the Bethe free energy on tree structures. Wainwright et al. (2002b) use the tree structure to obtain a more efficient implementation of loopy belief propagation, without however guaranteeing convergence. Wainwright et al. (2002a) show that particular convex combinations of convex Bethe free energies lead to convex bounds on the exact Helmholtz free energy (2). In these bounds, the overcounting numbers of the inner regions still follow the Moebius relation (4), but the overcounting numbers for the outer regions are smaller than or equal to 1. Constrained minimization of such a bound is very similar to constrained minimization of $F^{(3)}_{\text{convex}}$ and the algorithm used by Wainwright, Jaakkola, and Willsky (2003) is indeed closely related to Algorithm 1.

4. The procedure described by Yuille (2002) often even moves part of the convex terms to the concave side. This makes the (implicit) bound even worse and the corresponding algorithm slower. In the following we will stick to the more favorable interpretation of the CCCP algorithm that is based on the implicit bound (26).

5. Simulations

Intuitively, we would expect the algorithms based on the tightest bound to converge the fastest in terms of outer-loop iterations. However, with larger steps in the outer loop, we might need more inner-loop iterations to achieve convergence in the inner loop. The following simulations are designed to check this.

5.1 General Set-up

In the simulations we compare four different algorithms, each of them based on a different bound.

`just_convex` The tightest bound of the Kikuchi free energy that is just convex. Based on the ideas described in Section 4.4 and Appendix C.

`negative_to_zero` The bound obtained by setting all negative overcounting numbers to zero, as explained in Section 4.2.

`all_to_zero` The bound described in Section 4.3 that follows by setting all overcounting numbers, both negative and positive, to zero. In all models considered below, the overcounting numbers satisfy the conditions of Theorem 4.2, i.e., setting them to zero indeed yields a bound on the Kikuchi free energy. Note further that `all_to_zero` is equivalent to `negative_to_zero` for the Bethe free energy.

`cccp` The (rather favorable interpretation of the) bound implicit in Yuille's (2002) CCCP algorithm, as explained in Section 4.5.

Algorithm 1 is applied in the inner loop of all these algorithms: the only difference between them is the setting of the overcounting numbers $\tilde{c}_\beta$ implied by the bound. Each inner loop runs until a preset convergence criterion is met. Specifically, we end the inner loop when all inner region marginals change less then 10^{-4}. With this criterion all algorithms happened to converge, which probably would also have been the case with looser criteria. For example, Yuille (2002) reports that two inner-loop iterations were sufficient to obtain convergence.

In all simulations we report on the Kullback-Leibler (KL) divergence between exact and approximate marginals, either summed over all nodes or over a subset of nodes. Plots for the different error functions all look very much the same. The Kikuchi/Bethe free energy itself is somewhat less illustrative: when it is very close to its minimum, the marginals and thus KL divergence can still change considerably. We visualize the KL divergence both as a function of outer-loop iterations and as a function of floating point operations, where we count only the necessary operations involved in the inner-loop and outer-loop updates (i.e., not those involved in convergence checks, computing the KL divergence, and so on). In comparing the number of inner-loop iterations used by the different algorithms to meet the convergence criterion, we scale the outer-loop iterations relative to the outer-loop iterations of the `just_convex` algorithm. That is, for each number of outer-loop iterations used by an algorithm to reach a particular level of accuracy, we consider the corresponding number of outer-loop iterations used by the `just_convex` algorithm to reach the same level.

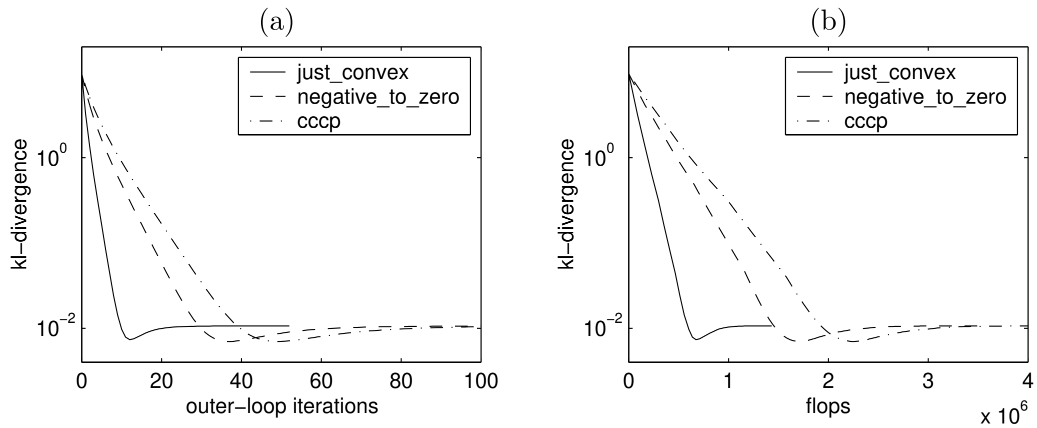


Figure 4: Bethe approximation on a 9×9 Boltzmann grid. Kullback-Leibler divergence between exact and approximate single-node marginals as a function of the outer-loop iterations (a) and floating point operations (b) for three different algorithms.

We have done simulations on quite a number of different problems and problem instances, involving both Markov random fields and Bayesian networks. The results shown here are exemplary and meant to illustrate the more general findings that we will summarize below.

5.2 Bethe Free Energy on a Boltzmann Grid

Our first set of simulations concerns the minimization of the Bethe free energy on a Boltzmann grid of 9×9 nodes with pairwise interactions of the form

$$\psi_{ij}(x_i, x_j) = \exp\left[w_{ij}(2x_i - 1)(2x_j - 1) + \frac{t_i}{n_i}(2x_i - 1) + \frac{t_j}{n_j}(2x_j - 1)\right] \tag{27}$$

where n_i is the number of neighbors of node i, i.e., 2 for a corner node, 3 for other nodes on the boundary, and 4 for nodes in the middle. Weights w_{ij} and biases t_i are drawn at random from a normal distribution with mean zero and standard deviation 0.5. In the Bethe approximation the outer regions are all pairs of neighboring nodes.

Figure 4 shows the summed KL divergence between exact and approximate single-node marginals as a function of the number of outer loop iterations (a) and as a function of the number of floating point operations (b) for the `just_convex`, `negative_to_zero`, and `cccp` algorithms. It can be seen that, as expected, the `just_convex` algorithms converges faster than the `negative_to_zero` algorithm, which itself converges faster than the `cccp` algorithm. The speed-up in terms of outer-loop iterations translates into an almost equivalent speed-up in terms of flops. Indeed, as can be seen in Figure 5(a), the number of inner-loop iterations required by the `just_convex` algorithm is just slightly higher than that of the other two algorithms.

The curves in Figure 4(a) can be mapped onto each other with a rough linear scaling of the number of outer-loop iterations. This is also suggested by the straight lines in

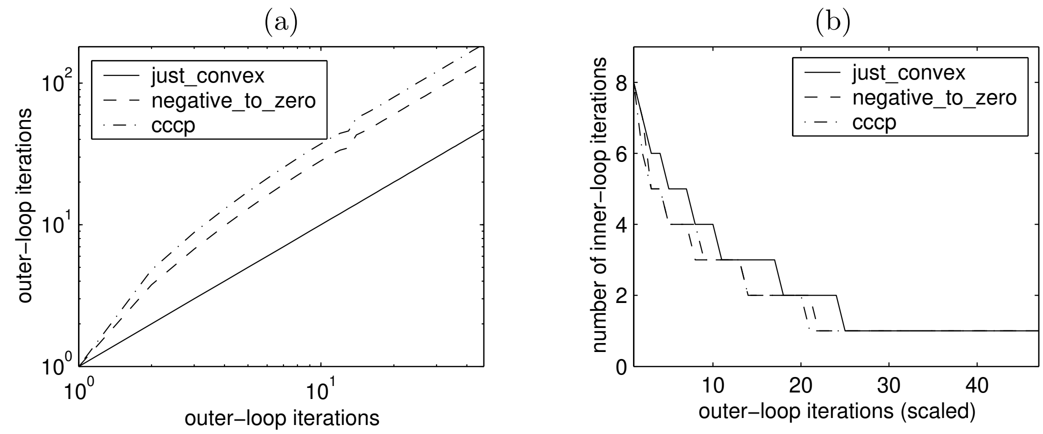


Figure 5: Bethe approximation on a 9×9 Boltzmann grid. (a) Outer loop iterations of the `just_convex` algorithm versus the corresponding outer-loop iterations of the other two algorithms. (b) Number of inner loop iterations needed to meet the convergence criterion as a function of the outer-loop iterations, scaled according to (a).

Figure 5(a). The slope of these lines relate to each other as 0.34, 1 (by definition), and 1.35 for `just_convex`, `negative_to_zero` and `cccp`, respectively (see also the convergence rates in Table 1). The following argumentation shows that there is a striking correspondence between these numbers and the respective bounds. The negative overcounting numbers for the Bethe free energy F_{Kikuchi} add up to $\sum_{\beta \in \mathcal{I}_-} c_\beta = -207$. For the respective convex bounds F_{convex}, these sums are $\sum_{\beta \in \mathcal{I}_-} \tilde{c}_\beta = -144$, 0, and 81. If we now translate these into the fraction of "negative overcounting mass" that is "bounded", i.e.,

$$\frac{\sum_{\beta \in \mathcal{I}_-} c_\beta - \sum_{\beta \in \mathcal{I}_-} \tilde{c}_\beta}{\sum_{\beta \in \mathcal{I}_-} c_\beta},$$

we obtain, respectively 0.30, 1 (by definition), and 1.39. That is, there appears to be an almost linear relationship between the tightness of the bound (here expressed in the fraction of concave entropy contributions that is bounded linearly) and the speed of convergence. We have noticed the same almost linear relationship in all other simulations involving a Bethe free energy (no positive overcounting numbers).

5.3 Kikuchi Free Energy on a Boltzmann Grid

Our second set of simulations is also on a 9×9 Boltzmann grid, where now the outer regions are chosen to be all squares of four neighboring nodes. Potentials are of the form (27) with weights and biases drawn from a normal distribution with standard deviation 4 and 0.5, respectively. Note that the size of the weights is much larger than in the previous set of simulations, to make the problem still a bit of a challenge for the Kikuchi approximation. With these weights, the Bethe approximation does very badly (summed Kullback-Leibler

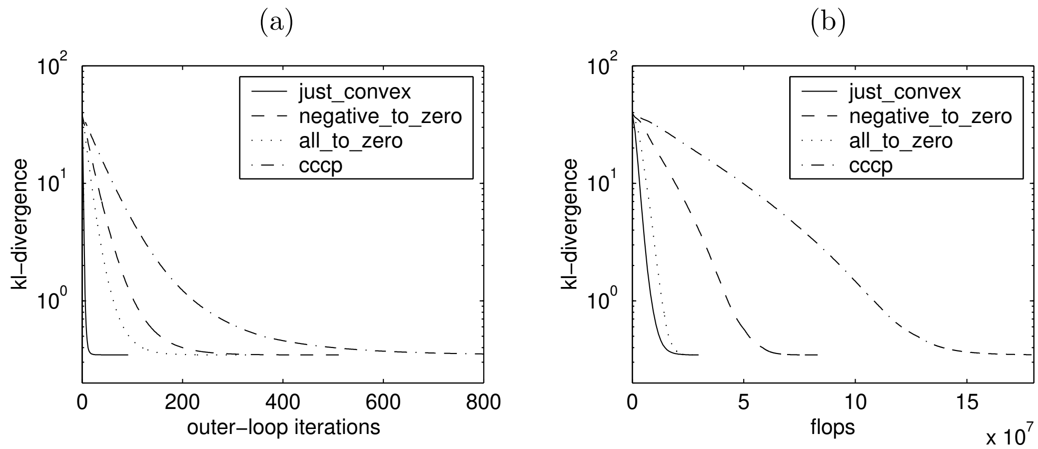


Figure 6: Kikuchi approximation on a 9×9 Boltzmann grid. Kullback-Leibler divergence between exact and approximate single-node marginals as a function of the outer-loop iterations (a) and floating point operations (b) for four different algorithms.

divergence larger than 10). Both for the Bethe and for the Kikuchi algorithm, the single-loop algorithm has convergence problems: for the Bethe approximation it typically gets stuck in a limit cycle and for the Kikuchi approximation it tends to diverge. In total there are $8 \times 8 = 64$ outer regions and $(8 \times 7) \times 2 = 122$ negative inner regions (all node pairs that correspond to intersections of the outer regions) and $7 \times 7 = 49$ positive inner regions (all single nodes that correspond to intersections of the node pairs).

Figure 6 shows the KL divergence between approximate and exact single-node marginals for the four different algorithms in terms of the outer-loop iterations (a) and floating point operations (b). It can be seen that the ordering in (a) is again as expected: the tighter the bound, the faster the algorithm. In terms of floating point operations, the `just_convex` and `all_to_zero` algorithm get much closer together.

Part of the explanation is given in Figure 7: the `just_convex` algorithm requires considerably more inner-loop iterations to meet the same convergence criterion. The other effect is that the `all_to_zero` algorithm in its inner loop only runs over the 112 negative inner regions instead of all 161 positive and negative inner regions. This makes that each inner-loop iteration of `all_to_zero` requires a factor 1.8 less floating point operations than an inner-loop iteration of the other three algorithms.

Here it is more difficult to find a quantitative relationship between the tightness of the bounds and the (asymptotic) convergence rates. One of the complications is that not only the negative, but also the positive overcounting numbers play a role. In any case, all algorithms still seem to converge linearly, with faster convergence rates for tighter bounds. These convergence rates, expressed as the time scale of the corresponding exponential decay ($\mathrm{KL}(t) - \mathrm{KL}(\infty) \propto \exp[-t/\tau]$, with t and τ in outer-loop iterations), are summarized in Table 1.

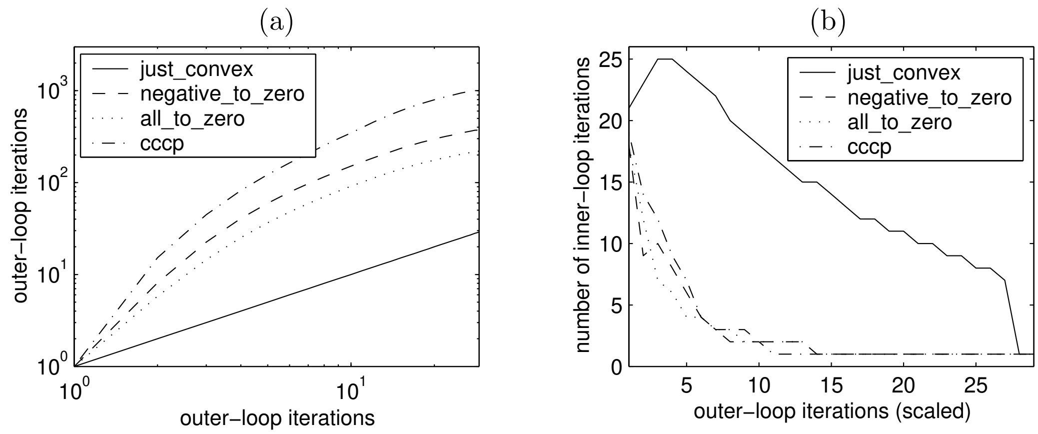


Figure 7: Kikuchi approximation on a 9×9 Boltzmann grid. (a) Outer loop iterations of the `just_convex` algorithm versus the corresponding outer-loop iterations of the other three algorithms. (b) Number of inner loop iterations needed to meet the convergence criterion as a function of the outer-loop iterations, scaled according to (a).

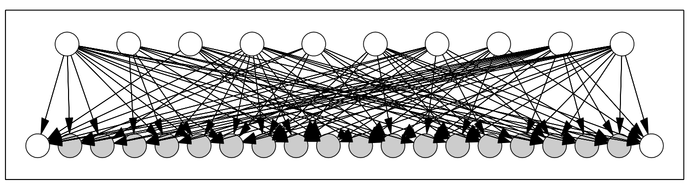


Figure 8: Graphical structure of the QMR-like network.

5.4 A QMR Network

Our third set of simulations concerns a QMR-like (Quick Medical Reference) Bayesian network (Heckerman, 1989; Jaakkola & Jordan, 1999): a bipartite graph with a layer of disease nodes and a layer of findings. The particular network used in these simulations has been generated with the Bayes Net Toolbox (Murphy, 2001). It contains 20 finding nodes, of which 18 are observed (positive), and 10 hidden disease nodes; see Figure 8. The diseases have Bernoulli probability distributions with a prior drawn at random between 0 and 0.01. The findings have noisy-or conditional probability distributions without leakage. Diseases and findings are linked randomly with probability 0.5. The absence of leakage, large amount of findings, and strong connectivity make this a relatively difficult inference problem. As outer regions we take the subsets implied by the conditional probability distribution, i.e., each outer region consists of a disease and all findings linked to it. Figure 9 gives the corresponding region graph.

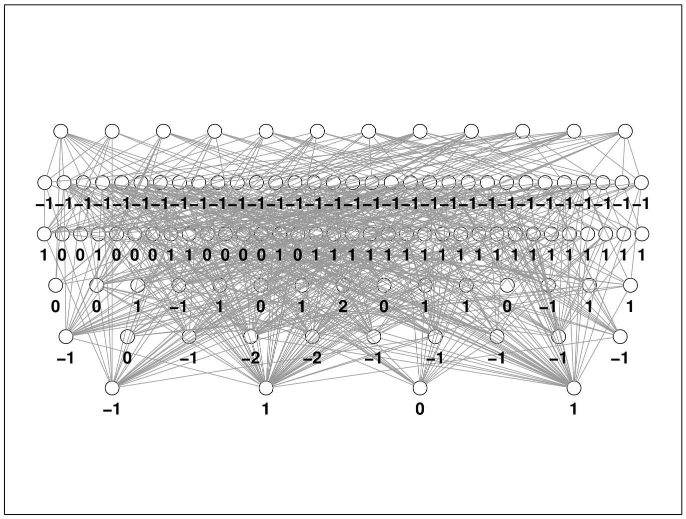


Figure 9: Region graph resulting from the QMR-like network.

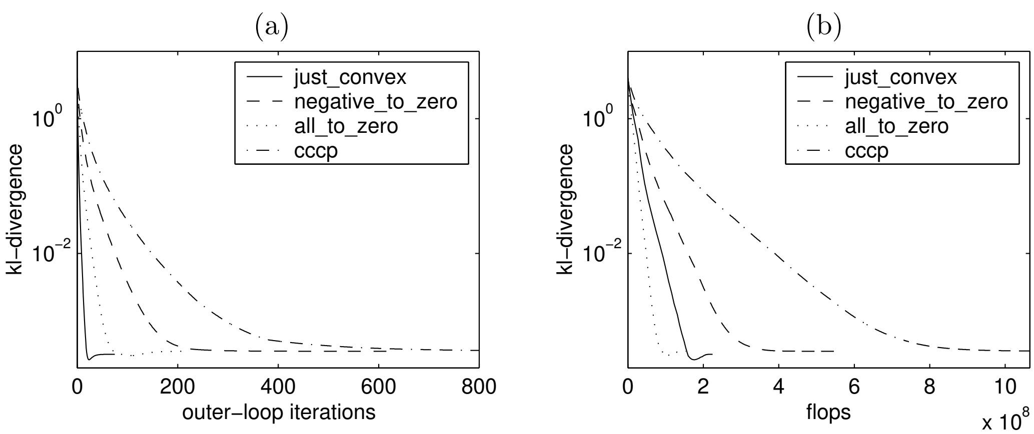


Figure 10: Kikuchi approximation on a QMR-like network. Kullback-Leibler divergence between exact and approximate single-node marginals as a function of the outer-loop iterations (a) and floating point operations (b) for four different algorithms.

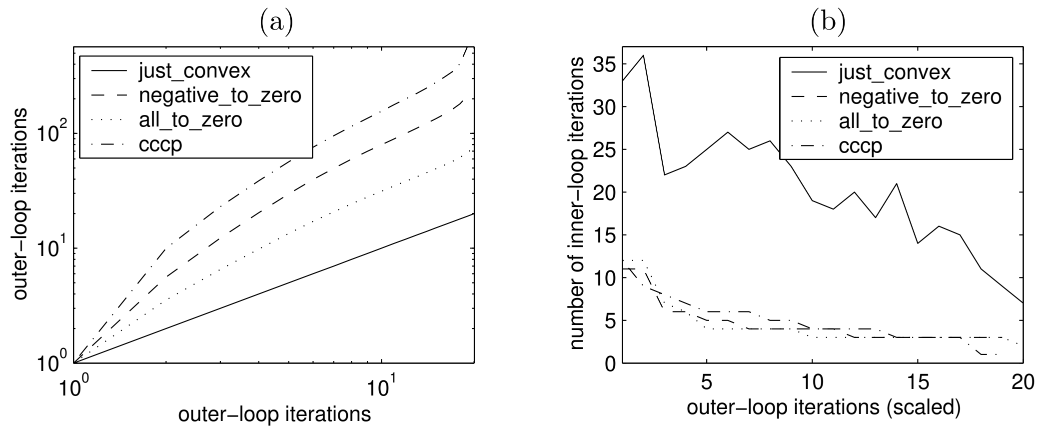


Figure 11: Kikuchi approximation on a QMR-like network. (a) Outer loop iterations of the `just_convex` algorithm versus the corresponding outer-loop iterations of the other three algorithms. (b) Number of inner loop iterations needed to meet the convergence criterion as a function of the outer-loop iterations, scaled according to (a).

	Bethe			Kikuchi			QMR		
	-	+	τ	-	+	τ	-	+	τ
original	-207	0	-	-112	49	-	-54	35	-
`just_convex`	-144	0	3.8	-64	1	11	-34	15	6
`negative_to_zero`	0	0	11.3	0	49	41	0	35	67
`all_to_zero`	0	0	11.3	0	0	29	0	0	17
`cccp`	81	0	15.3	112	49	153	52	35	166

Table 1: Summary of asymptotic convergence (τ is the time constant, with time in outer-loop iterations, in the exponential decay) and sums of negative and positive overcounting numbers in the original Kikuchi/Bethe free energy and the convex bounds used by the different algorithms.

The results can be found in Figure 10 and 11. They are comparable with those for the Kikuchi approximation on the Boltzmann grid. Also here the single-loop algorithm fails to converge. The `just_convex` algorithm converges much faster than the other three algorithms, but requires more inner-loop iterations and is less efficient than the `all_to_zero` algorithm, which makes the latter preferable in terms of floating point operations. However, it is relatively straightforward to speed-up the `just_convex` algorithm. First, we probably do not need that many inner-loop iterations for the outer loop to converge properly. And secondly, where we now bound part of each entropy contribution, a more efficient choice would have as many zero overcounting numbers as possible.

5.5 General Findings

Here we summarize some of the points that have been illustrated above and that we have encountered in many other simulations as well.

- The tighter the (convex) bound used in the inner loop, the faster the convergence in terms of outer-loop iterations.
- The number of outer-loop iterations needed to meet a prespecified convergence criterion tends to decrease with a looser bound, but never nearly enough to compensate for the slower convergence in the outer loop.
- In fact, we have only observed a strong dependency between this number of inner-loop iterations and the tightness of the bound if the bound is just convex and the problem is "hard" in the sense that a single-loop algorithm would fail to converge.
- In terms of floating point operations, a looser bound that sets all overcounting numbers in the inner loop to zero, can beat a tighter bound with negative overcounting numbers: the slower convergence in terms of outer-loop iterations is compensated by a more efficient inner loop.

Pelizzola (2005) tests several convergent algorithms on Kikuchi approximations of problems in statistical physics and reports similar findings. Also in this study, the `just_convex` algorithm, described for the first time by Heskes, Albers, and Kappen (2003), clearly outperforms all competitors.

6. Discussion

This article is based on the perspective that we are interested in *minima* of the Kikuchi free energy under appropriate constraints. Finding such a minimum then becomes a possibly non-convex constrained minimization problem. Here, as well as in other studies, the approach has been to solve this non-convex problem through sequential constrained minimization of convex bounds on the Kikuchi free energy. On the presumption that tighter bounds yield faster algorithms, we have worked out several ideas to construct tight convex bounds. The simulation results in this article as well as those obtained by Pelizzola (2005) clearly validate this presumption and show that the speed-ups can be very significant. Heskes, Zoeter, and Wiegerinck (2004) apply these bounds for (approximate) parameter learning in directed graphical models.

The double-loop algorithms considered in this article are all based on convex bounds of the Kikuchi free energy. In principle, this is not necessary: our only concern is that the inner-loop algorithm converges and this might well be the case for tighter bounds. One practical solution is to simply choose a (tight) bound on the Kikuchi free, check whether the inner-loop algorithm does converge, and restart with a looser bound if not. Alternatively, we can construct tighter bounds making use of conditions for guaranteed convergence of belief propagation such as those derived by Tatikonda and Jordan (2002), Heskes (2004), Ihler et al. (2005) for the Bethe approximation.

It has been suggested that non-convergence of single-loop generalized/loopy belief propagation by itself is an indication that the Kikuchi/Bethe approximation is inaccurate. The results in Section 5.3 and 5.4 show that this need not always be the case. Apparently, there does exist a "middle range" of problems where the Kikuchi free energy is not easy to minimize, but does yield decent approximations. It is on these problems that the algorithms described in this article are useful.

Acknowledgments

The author would like to thank Wim Wiegerinck, Onno Zoeter, Kees Albers, and Bert Kappen for fruitful discussions and the anonymous reviewers for their constructive comments. This work has been supported in part by the Dutch Technology Foundation STW.

Appendix A: Convexity of the Difference between Two Entropies

This appendix treats two lemmas on the convexity of the difference between two entropies. The first one is used in the proof of Theorem 3.1. A similar lemma is used by McEliece and Yildirim (2003).

Lemma A.1. *The difference between two entropies*

$$\begin{aligned}\Delta_{\gamma\beta}(q_\gamma) &\equiv \sum_{\mathbf{x}_\gamma} q_\gamma(\mathbf{x}_\gamma)\log q_\gamma(\mathbf{x}_\gamma) - \sum_{\mathbf{x}_\beta} q_\gamma(\mathbf{x}_\beta)\log q_\gamma(\mathbf{x}_\beta)\\ &= \sum_{\mathbf{x}_\beta} q_\gamma(\mathbf{x}_\beta)\left[\sum_{\mathbf{x}_{\gamma\backslash\beta}} q_\gamma(\mathbf{x}_{\gamma\backslash\beta}|\mathbf{x}_\beta)\log q_\gamma(\mathbf{x}_{\gamma\backslash\beta}|\mathbf{x}_\beta)\right]\end{aligned}$$

is convex in q_γ.

Proof We take a step backwards and write $\Delta_{\gamma\beta}$ out as

$$\Delta_{\gamma\beta}(q_\gamma) = \sum_{\mathbf{x}_\gamma} q_\gamma(\mathbf{x}_\gamma)\log\left[\frac{q_\gamma(\mathbf{x}_\gamma)}{\sum_{\mathbf{x}'_{\gamma\backslash\beta}} q_\gamma(\mathbf{x}'_\gamma)}\right].$$

When taking derivatives, we best interpret the table q_γ, specifying the value $q_\gamma(\mathbf{x}_\gamma)$ for each possible realization $\mathbf{x}_\gamma$, as a vector with $\mathbf{x}_\gamma$ playing the role of an index. Taking second derivatives, we then obtain

$$H_{\mathbf{x}_\gamma,\mathbf{x}'_\gamma}(q_\gamma) \equiv \frac{\partial^2 \Delta_{\gamma\beta}(q_\gamma)}{\partial q_\gamma(\mathbf{x}_\gamma)\partial q_\gamma(\mathbf{x}'_\gamma)} = \frac{1}{q_\gamma(\mathbf{x}_\gamma)} I_{\mathbf{x}_\gamma,\mathbf{x}'_\gamma} - \frac{1}{q_\gamma(\mathbf{x}_\beta)} I_{\mathbf{x}_\beta,\mathbf{x}'_\beta}\,.$$

with $I_{\mathbf{x},\mathbf{x}'} \equiv 1$ if all elements of $\mathbf{x}$ and $\mathbf{x}'$ are equal and zero otherwise.

Next we would like to show that this matrix is positive semi-definite, i.e., that for all tables $\tilde{q}$, again to be interpreted as vectors with indices $\mathbf{x}_\gamma$,

$$
\begin{aligned}
0 &\leq \sum_{\mathbf{x}_\gamma, \mathbf{x}'_\gamma} \tilde{q}(\mathbf{x}_\gamma) H_{\mathbf{x}_\gamma, \mathbf{x}'_\gamma}(q_\gamma) \tilde{q}(\mathbf{x}'_\gamma) = \sum_{\mathbf{x}_\gamma} \frac{\tilde{q}^2(\mathbf{x}_\gamma)}{q_\gamma(\mathbf{x}_\gamma)} - \sum_{\mathbf{x}_\gamma, \mathbf{x}'_\gamma} \frac{\tilde{q}(\mathbf{x}_\gamma)\tilde{q}(\mathbf{x}'_\gamma)}{q_\gamma(\mathbf{x}_\beta)} I_{\mathbf{x}_\beta, \mathbf{x}'_\beta} \\
&= \sum_{\mathbf{x}_\beta} \left\{ \sum_{\mathbf{x}_{\gamma\backslash\beta}} \frac{\tilde{q}^2(\mathbf{x}_{\gamma\backslash\beta}, \mathbf{x}_\beta)}{q_\gamma(\mathbf{x}_{\gamma\backslash\beta}, \mathbf{x}_\beta)} - \sum_{\mathbf{x}_{\gamma\backslash\beta}, \mathbf{x}'_{\gamma\backslash\beta}} \frac{\tilde{q}(\mathbf{x}_{\gamma\backslash\beta}, \mathbf{x}_\beta)\tilde{q}(\mathbf{x}'_{\gamma\backslash\beta}, \mathbf{x}_\beta)}{q_\gamma(\mathbf{x}_\beta)} \right\} \\
&= \sum_{\mathbf{x}_\beta} \left\{ \sum_{\mathbf{x}_{\gamma\backslash\beta}} \frac{\tilde{q}^2(\mathbf{x}_{\gamma\backslash\beta}, \mathbf{x}_\beta)}{q_\gamma(\mathbf{x}_{\gamma\backslash\beta}, \mathbf{x}_\beta)} - \frac{\left[\sum_{\mathbf{x}_{\gamma\backslash\beta}} \tilde{q}(\mathbf{x}_{\gamma\backslash\beta}, \mathbf{x}_\beta)\right]^2}{\sum_{\mathbf{x}_{\gamma\backslash\beta}} q_\gamma(\mathbf{x}_{\gamma\backslash\beta}, \mathbf{x}_\beta)} \right\} .
\end{aligned}
$$

From Cauchy's inequality,

$$
\sum_k a_k^2 \sum_k b_k^2 \geq \left[\sum_k a_k b_k \right]^2 ,
$$

it follows that the term between braces is indeed semi-positive for each realization of $\mathbf{x}_\beta$. To see this, we make the substitutions $\mathbf{x}_{\gamma\backslash\beta} \Rightarrow k$, $\tilde{q}(\mathbf{x}_{\gamma\backslash\beta}, \mathbf{x}_\beta)/\sqrt{q_\gamma(\mathbf{x}_{\gamma\backslash\beta}, \mathbf{x}_\beta)} \Rightarrow a_k$, and $\sqrt{q_\gamma(\mathbf{x}_{\gamma\backslash\beta}, \mathbf{x}_\beta)} \Rightarrow b_k$ to find

$$
\{\ldots\} \Rightarrow \sum_k a_k^2 - \frac{\left[\sum_k a_k b_k\right]^2}{\sum_k b_k^2} \geq 0 . \quad \square
$$

The following related lemma is used in Appendix B.

Lemma A.2. *The difference between two entropies*

$$
\Delta_{\gamma\beta}(q_\gamma, q_\beta) \equiv \sum_{\mathbf{x}_\gamma} q_\gamma(\mathbf{x}_\gamma) \log q_\gamma(\mathbf{x}_\gamma) - \sum_{\mathbf{x}_\beta} q_\gamma(\mathbf{x}_\beta) \log q_\beta(\mathbf{x}_\beta)
$$

is convex in $\{q_\gamma, q_\beta\}$.

Proof The Hessian matrix has components

$$
\begin{aligned}
H_{\mathbf{x}_\gamma, \mathbf{x}'_\gamma} &\equiv \frac{\partial^2 \Delta_{\gamma\beta}(q_\gamma)}{\partial q_\gamma(\mathbf{x}_\gamma) \partial q_\gamma(\mathbf{x}'_\gamma)} = \frac{1}{q_\gamma(\mathbf{x}_\gamma)} I_{\mathbf{x}_\gamma, \mathbf{x}'_\gamma} \\
H_{\mathbf{x}_\gamma, \mathbf{x}'_\beta} &\equiv \frac{\partial^2 \Delta_{\gamma\beta}(q_\gamma)}{\partial q_\gamma(\mathbf{x}_\gamma) \partial q_\beta(\mathbf{x}'_\beta)} = -\frac{1}{q_\beta(\mathbf{x}_\beta)} I_{\mathbf{x}_\beta, \mathbf{x}'_\beta} \\
H_{\mathbf{x}_\beta, \mathbf{x}'_\beta} &\equiv \frac{\partial^2 \Delta_{\gamma\beta}(q_\gamma)}{\partial q_\beta(\mathbf{x}_\beta) \partial q_\beta(\mathbf{x}'_\beta)} = \frac{q_\gamma(\mathbf{x}_\beta)}{q_\beta^2(\mathbf{x}_\beta)} I_{\mathbf{x}_\beta, \mathbf{x}'_\beta} .
\end{aligned}
$$

Convexity requires that for any $\tilde{q} = (\tilde{q}_\gamma(\mathbf{x}_\gamma), \tilde{q}_\beta(\mathbf{x}_\beta))$,

$$
\begin{aligned}
0 &\leq \begin{pmatrix} \tilde{q}_\gamma(\mathbf{x}_\gamma) & \tilde{q}_\beta(\mathbf{x}_\beta) \end{pmatrix} \begin{pmatrix} H_{\mathbf{x}_\gamma,\mathbf{x}'_\gamma} & H_{\mathbf{x}_\gamma,\mathbf{x}'_\beta} \\ H_{\mathbf{x}_\beta,\mathbf{x}'_\gamma} & H_{\mathbf{x}_\beta,\mathbf{x}'_\beta} \end{pmatrix} \begin{pmatrix} \tilde{q}_\gamma(\mathbf{x}'_\gamma) \\ \tilde{q}_\beta(\mathbf{x}'_\beta) \end{pmatrix} \\
&= \sum_{\mathbf{x}_\gamma} \frac{\tilde{q}^2_\gamma(\mathbf{x}_\gamma)}{q_\gamma(\mathbf{x}_\gamma)} - 2\sum_{\mathbf{x}_\gamma} \frac{\tilde{q}_\gamma(\mathbf{x}_\gamma)\tilde{q}_\beta(\mathbf{x}_\beta)}{q_\beta(\mathbf{x}_\beta)} + \sum_{\mathbf{x}_\beta} \frac{q_\gamma(\mathbf{x}_\beta)\tilde{q}^2_\beta(\mathbf{x}_\beta)}{q^2_\beta(\mathbf{x}_\beta)} \\
&= \sum_{\mathbf{x}_\gamma} q_\gamma(\mathbf{x}_\gamma) \left[\frac{\tilde{q}_\gamma(\mathbf{x}_\gamma)}{q_\gamma(\mathbf{x}_\gamma)} - \frac{\tilde{q}_\beta(\mathbf{x}_\beta)}{q_\beta(\mathbf{x}_\beta)}\right]^2 . \qquad \square
\end{aligned}
$$

Appendix B: Minimizing a Convex Kikuchi Free Energy

In this appendix, we derive Algorithm 1 for minimizing a convex Kikuchi free energy under appropriate linear constraints. To simplify notation, we will use the convention that α runs over outer regions, and β over inner regions.

First, we note that in principle it is not necessary to explicitly take into account all constraints (6), since some constraints are implied by others. Obviously, the constraint between two inner region marginals,

$$q_{\beta'}(\mathbf{x}_\beta) = q_\beta(\mathbf{x}_\beta) \ \text{ for some } \beta' \supset \beta \,,$$

is implied by corresponding constraints between the inner region marginals and an outer region subsuming both inner regions,

$$q_\alpha(\mathbf{x}_{\beta'}) = q_{\beta'}(\mathbf{x}_{\beta'}) \ \text{ and } \ q_\alpha(\mathbf{x}_\beta) = q_\beta(\mathbf{x}_\beta) \ \text{ for some } \alpha \supset \beta \supset \beta'.$$

That is, we do not have to take into account constraints between inner regions and other inner regions. Similarly, normalization constraints on outer region pseudo-marginals follow from normalization constraints on the inner region pseudo-marginals. So, a sufficient set of constraints is

$$
\begin{aligned}
& q_\alpha(\mathbf{x}_\beta) = q_\beta(\mathbf{x}_\beta) \ \text{ with } \ q_\alpha(\mathbf{x}_\beta) = \sum_{\mathbf{x}_{\alpha\backslash\beta}} q_\alpha(\mathbf{x}_\alpha) \qquad \forall_{\alpha\supset\beta} \\
& \sum_{\mathbf{x}_\beta} q_\beta(\mathbf{x}_\beta) = 1 \qquad \forall_\beta \,.
\end{aligned}
$$

Introducing Lagrange multipliers $\lambda_{\alpha\beta}(\mathbf{x}_\beta)$ and λ_β for the corresponding constraints, we obtain the Lagrangian

$$
\begin{aligned}
L(\mathbf{q}, \boldsymbol{\lambda}) \ = \ & \sum_\alpha \sum_{\mathbf{x}_\alpha} q_\alpha(\mathbf{x}_\alpha) \log\left[\frac{q_\alpha(\mathbf{x}_\alpha)}{\psi_\alpha(\mathbf{x}_\alpha)}\right] + \sum_\beta c_\beta \sum_{\mathbf{x}_\beta} q_\beta(\mathbf{x}_\beta) \log q_\beta(\mathbf{x}_\beta) \\
& + \sum_\beta \sum_{\alpha\supset\beta} \sum_{\mathbf{x}_\beta} \lambda_{\alpha\beta}(\mathbf{x}_\beta) \left[q_\beta(\mathbf{x}_\beta) - \sum_{\mathbf{x}_{\alpha\backslash\beta}} q_\alpha(\mathbf{x}_\alpha)\right] + \sum_\beta \lambda_\beta \left[1 - \sum_{\mathbf{x}_\beta} q_\beta(\mathbf{x}_\beta)\right] . \quad \text{(B-1)}
\end{aligned}
$$

Convex Independent of the Constraints

Let us first consider the case that all overcounting numbers c_β are strictly positive ($c_\beta > 0$). Then, the Lagrangian is not just convex over the set of constraints, but convex in $\mathbf{q}$ independent of the constraints. Minimization of the Lagrangian with respect to these pseudo-marginals follows by setting its derivatives to zero, yielding

$$
\begin{aligned}
q^*_\alpha(\mathbf{x}_\alpha) &= \psi_\alpha(\mathbf{x}_\alpha) e^{-1} \prod_{\beta\subset\alpha} e^{\lambda_{\alpha\beta}(\mathbf{x}_\beta)} \quad \forall_\alpha && \text{(B-2)} \\
q^*_\beta(\mathbf{x}_\beta) &= e^{\lambda_\beta/c_\beta - 1} \prod_{\alpha\supset\beta} e^{-\lambda_{\alpha\beta}(\mathbf{x}_\beta)/c_\beta} \quad \forall_\beta\,, && \text{(B-3)}
\end{aligned}
$$

where here and in the following it should be noted that q^*_α and q^*_β are functions of the Lagrange multipliers $\boldsymbol{\lambda}$. Substituting this solution back into the Lagrangian, we obtain the dual

$$
L^*(\boldsymbol{\lambda}) \equiv L(\mathbf{q}^*(\boldsymbol{\lambda}), \boldsymbol{\lambda}) = \sum_\beta \lambda_\beta - \sum_\alpha \sum_{\mathbf{x}_\alpha} q_\alpha(\mathbf{x}_\alpha) - \sum_\beta c_\beta \sum_{\mathbf{x}_\beta} q_\beta(\mathbf{x}_\beta)\,. \tag{B-4}
$$

Now, consider optimizing $L^*(\boldsymbol{\lambda})$ with respect to the subset of the components corresponding to the inner region β, collected in $\boldsymbol{\lambda}_\beta \equiv (\lambda_\beta, \lambda_{\alpha\beta}(\mathbf{x}_\beta) \;\; \forall_{\alpha\supset\beta,\mathbf{x}_\beta})$, keeping all other $\boldsymbol{\lambda}_{\beta'}$ for $\beta' \neq \beta$ fixed. Because of the concavity of the dual $L^*(\boldsymbol{\lambda})$, we can find the maximum in the direction $\boldsymbol{\lambda}_\beta$ by setting the corresponding derivatives to zero. This yields

$$
\begin{aligned}
\left.\frac{\partial L^*(\boldsymbol{\lambda})}{\partial \lambda_{\alpha\beta}(\mathbf{x}_\beta)}\right|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^{\text{new}}} &= q^{\text{new}}_\beta(\mathbf{x}_\beta) - q^{\text{new}}_\alpha(\mathbf{x}_\beta) = 0 \quad \forall_{\mathbf{x}_\beta;\alpha\supset\beta} \\
\left.\frac{\partial L^*(\boldsymbol{\lambda})}{\partial \lambda_\beta}\right|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^{\text{new}}} &= 1 - \sum_{\mathbf{x}_\beta} q^{\text{new}}_\beta(\mathbf{x}_\beta) = 0\,, && \text{(B-5)}
\end{aligned}
$$

where q^{new} refers to the solution (B-2) and (B-3) with $\lambda_{\alpha\beta}(\mathbf{x}_\beta)$ replaced by $\lambda^{\text{new}}_{\alpha\beta}(\mathbf{x}_\beta)$ and λ_β by $\lambda^{\text{new}}_\beta$. Since from (B-2)

$$
q^{\text{new}}_\alpha(\mathbf{x}_\beta) = \frac{e^{\lambda^{\text{new}}_{\alpha\beta}(\mathbf{x}_\beta)}}{e^{\lambda_{\alpha\beta}(\mathbf{x}_\beta)}} q_\alpha(\mathbf{x}_\beta)\,,
$$

the solution for $\lambda^{\text{new}}_{\alpha\beta}(\mathbf{x}_\beta)$ must obey

$$
\lambda^{\text{new}}_{\alpha\beta}(\mathbf{x}_\beta) = -\log q_\alpha(\mathbf{x}_\beta) + \lambda_{\alpha\beta}(\mathbf{x}_\beta) + \log q^{\text{new}}_\beta(\mathbf{x}_\beta)\,,
$$

where we still have to solve for $q^{\text{new}}_\beta(\mathbf{x}_\beta)$. Summing this expression over all $\alpha \supset \beta$, substituting (B-3), and solving for $q^{\text{new}}_\beta(\mathbf{x}_\beta)$ we get

$$
\log q^{\text{new}}_\beta(\mathbf{x}_\beta) = \frac{1}{n_\beta + c_\beta} \sum_{\alpha\supset\beta} \left[\log q_\alpha(\mathbf{x}_\beta) - \lambda_{\alpha\beta}(\mathbf{x}_\beta)\right] + \frac{1}{n_\beta + c_\beta}(\lambda^{\text{new}}_\beta - c_\beta)\,.
$$

Now, we obtain exactly the updates in Algorithm 1 if we define

$$
\mu_{\beta\to\alpha}(\mathbf{x}_\beta) = e^{\lambda_{\alpha\beta}(\mathbf{x}_\beta)} \quad \text{and} \quad \mu_{\alpha\to\beta}(\mathbf{x}_\beta) = q_\alpha(\mathbf{x}_\beta) e^{-\lambda_{\alpha\beta}(\mathbf{x}_\beta)}\,,
$$

and properly normalize $q_\beta(\mathbf{x}_\beta)$, as in line 7. The normalization of $q_\alpha(\mathbf{x}_\alpha)$ in line 10 is then in fact unnecessary, since by construction the updates ensure that $q_\alpha(\mathbf{x}_\beta) = q_\beta(\mathbf{x}_\beta)$ with $Z_\alpha = 1$.

The bottom line is that with the particular ordering in Algorithm 1 the joint update of all messages for a particular subset β can be interpreted as doing coordinate-wise gradient ascent in the dual $L^*(\boldsymbol{\lambda})$, updating the Lagrange multipliers λ_β and $\lambda_{\alpha\beta}(\mathbf{x}_\beta)$ for a particular β and all $\alpha \supset \beta$ at the same time. Therefore Algorithm 1 is guaranteed to converge to the unique maximum in the case of all positive overcounting numbers c_β.

Convex over the Set of Constraints

Next, let us consider the more general case in which (some of) the overcounting numbers are negative, but such that the Kikuchi free energy is still convex over the set of constraints. We consider the case in which all inner region overcounting numbers are negative[5]. We will show that, with sufficient damping of the updates, Algorithm 1 is still guaranteed to converge to the unique minimum of the Kikuchi free energy over the set of constraints.

Note that direct application of the above argumentation fails, because the solution (B-3) for $q_\beta(\mathbf{x}_\beta)$ with negative c_β corresponds to a *maximum* rather than a minimum. Consequently, the dual $L^*(\boldsymbol{\lambda})$ in (B-4) need not be concave. The updates in Algorithm 1 that follow by setting derivatives to zero can be interpreted as fixed-point iterations, not as coordinate ascent in $L^*(\boldsymbol{\lambda})$. Still, in practice they do seem to work just fine and indeed without always increasing $L^*(\boldsymbol{\lambda})$. In the following we will explain why: we will argue that the updates of Algorithm 1 do not correspond to coordinate ascent, but rather to something like coordinate descent-ascent on a convex-concave saddle function. With sufficient damping, such an algorithm will converge to the unique saddle point, which then corresponds to the minimum of the Kikuchi free energy over the set of constraints.

Convexity over the set of constraints implies, according to Theorem 3.1, that there exists a matrix $A_{\alpha\beta}$ such that $\sum_\alpha A_{\alpha\beta} = |c_\beta|$ and $\sum_\beta A_{\alpha\beta} \leq 1$. Using $q_\beta(\mathbf{x}_\beta) = q_\alpha(\mathbf{x}_\beta)$, we replace the Lagrangian (B-1) by

$$\begin{aligned}\tilde{L}(\mathbf{q}, \boldsymbol{\lambda}) = {} & \sum_\alpha \sum_{\mathbf{x}_\alpha} q_\alpha(\mathbf{x}_\alpha) \log\left[\frac{q_\alpha(\mathbf{x}_\alpha)}{\psi_\alpha(\mathbf{x}_\alpha)}\right] - \sum_\beta \sum_{\alpha \supset \beta} A_{\alpha\beta} \sum_{\mathbf{x}_\beta} q_\alpha(\mathbf{x}_\beta) \log q_\beta(\mathbf{x}_\beta) \\ & + \sum_\beta \sum_{\alpha \supset \beta} \sum_{\mathbf{x}_\beta} \lambda_{\alpha\beta}(\mathbf{x}_\beta) \left[\frac{1}{n_\beta} \sum_{\alpha' \supset \beta} q_{\alpha'}(\mathbf{x}_\beta) - \sum_{\mathbf{x}_{\alpha \setminus \beta}} q_\alpha(\mathbf{x}_\alpha)\right] + \sum_\beta \lambda_\beta \left[1 - \sum_{\mathbf{x}_\beta} q_\beta(\mathbf{x}_\beta)\right] . \qquad \text{(B-6)}\end{aligned}$$

Now since, from Lemma A.2 in Appendix A,

$$\sum_{\mathbf{x}_\alpha} q_\alpha(\mathbf{x}_\alpha) \log q_\alpha(\mathbf{x}_\alpha) - \sum_{\mathbf{x}_\beta} q_\alpha(\mathbf{x}_\beta) \log q_\beta(\mathbf{x}_\beta)$$

is convex in $\{q_\alpha(\mathbf{x}_\alpha), q_\beta(\mathbf{x}_\beta)\}$, the Lagrangian (B-6) is indeed convex in $\mathbf{q}$ *independent* of the constraints. Thus we could apply the same argumentation as above: find the minimum of

5. Our argumentation does not hold if some of the negative inner region entropy contributions have to be compensated by positive inner region subset entropy contributions to prove convexity of the Kikuchi free energy. In that case, we might need a slightly different algorithm to guarantee convergence.

the convex Lagrangian with respect to $\mathbf{q}$, substitute the corresponding solution $\mathbf{q}^*(\boldsymbol{\lambda})$ back into the Lagrangian to obtain the concave dual $\tilde{L}^*(\boldsymbol{\lambda})$, and maximize this dual with respect to $\boldsymbol{\lambda}$. The problem is that we do not have a closed-form expression for the optimal $\mathbf{q}^*(\boldsymbol{\lambda})$ and thus also no closed-form expression for the dual $\tilde{L}^*(\boldsymbol{\lambda})$, which makes this procedure rather awkward.

Instead, we distinguish between the outer region marginals, collected in $\mathbf{q}_{\mathcal{O}}$, and the inner region marginals, collected in $\mathbf{q}_{\mathcal{I}}$. Having rewritten the consistency constraint in terms of outer region marginals alone, we only replace the constrained minimization with respect to $\mathbf{q}_{\mathcal{O}}$ by unconstrained maximization with respect to corresponding Lagrange multipliers $\boldsymbol{\lambda}_{\mathcal{O}}$, leaving the minimization with respect to $\mathbf{q}_{\mathcal{I}}$ under the normalization constraint as is. This gives us a saddle-point problem of the type $\min_{\mathbf{q}_{\mathcal{I}}} \max_{\boldsymbol{\lambda}_{\mathcal{O}}}$. Even without explicitly writing out the equations, we can tell that maximization with respect to $\lambda_{\alpha\beta}$ for a particular β and all $\alpha \supset \beta$ corresponds to finding $\lambda_{\alpha\beta}$ such that

$$q_\alpha^{\text{new}}(\mathbf{x}_\beta) = q_{\alpha'}^{\text{new}}(\mathbf{x}_\beta) \quad \forall_{\alpha,\alpha' \supset \beta} \, .$$

Then, minimization with respect to q_β given fixed $q_\alpha^{\text{new}}(\mathbf{x}_\beta)$ immediately yields

$$q_\beta^{\text{new}}(\mathbf{x}_\beta) \propto \sum_{\alpha \supset \beta} A_{\alpha\beta} q_\alpha^{\text{new}}(\mathbf{x}_\beta) \, ,$$

properly normalized to sum to 1. This is exactly what the updates for a particular inner region β in Algorithm 1 amount to: they yield the unique maximum with respect to $\lambda_{\alpha\beta}$ and minimum with respect to q_β, while keeping all other $\lambda_{\alpha'\beta'}$ and $q_{\beta'}$ for $\beta' \neq \beta$ fixed.

Such a "coordinate descent-ascent procedure" works fine if the saddle function is convex in the minimizing parameter and concave in the maximizing parameter (e.g., Seung, Richardson, Lagarias, & Hopfield, 1998). The concavity in $\boldsymbol{\lambda}$ is immediate, the convexity in $\mathbf{q}_{\mathcal{I}}$ follows from the convexity of the Lagrangian (B-6) in $\mathbf{q} = (\mathbf{q}_{\mathcal{O}}, \mathbf{q}_{\mathcal{I}})$: minimizing an overall convex function over some of its parameters, here $\mathbf{q}_{\mathcal{O}}$, yields a convex function over its remaining parameters, $\mathbf{q}_{\mathcal{I}}$. Technically, convergence to the unique solution of the saddle-point problem can be proven through the construction of a Lyapunov function that decreases under infinitesimal updates of the parameters in the descent and ascent direction to zero at the unique saddle point (Seung et al., 1998). Convergence can be guaranteed for sufficiently damped updates, not the "full" ones in Algorithm 1. Empirically the full updates, that correspond to full maximization and minimization for one inner region β before moving on the next one, work fine in most cases, but occasionally indeed require a little damping. Wainwright et al. (2003) successfully apply damping to a very similar algorithm in an attempt to minimize a convexified Bethe free energy.

Appendix C: Constructing a Tight Convex Bound

In this appendix, we describe a procedure for constructing a tight convex bound F_{convex} of the Kikuchi free energy F_{Kikuchi}. It combines ideas from Section 4.3 and 4.4. That is, we first convexify the Kikuchi free energy, bounding as little concave contributions from negative inner regions as possible. Next, in the terms that we have to bound anyways, we try to incorporate as many convex contributions as we can. This leads to the following procedure.

- Consider minus the entropy

$$-S = -\left\{ \sum_{\alpha \in \mathcal{O}} S_\alpha + \sum_{\beta \in \mathcal{I}_-} c_\beta S_\beta + \sum_{\gamma \in \mathcal{I}_+} c_\gamma S_\gamma \right\} ,$$

  and choose $\tilde{c}_\beta \geq c_\beta$ for $\beta \in \mathcal{I}_-$ such that the first term in

$$-S = -\left\{ \sum_{\alpha} S_\alpha + \sum_{\beta \in \mathcal{I}_-} \tilde{c}_\beta S_\beta + \sum_{\gamma \in \mathcal{I}_+} c_\gamma S_\gamma \right\} - \left\{ \sum_{\beta \in \mathcal{I}_-} (c_\beta - \tilde{c}_\beta) S_\beta \right\} ,$$

  is (just) convex.

- With A the corresponding allocation matrix of Theorem 3.1, define the "used resources"

$$\hat{c}_\gamma \equiv \sum_{\beta \in \mathcal{I}_-} A_{\gamma\beta} |\tilde{c}_\beta| \leq c_\gamma ,$$

  and rewrite

$$\begin{aligned} -S &= -\left\{ \sum_{\alpha} S_\alpha + \sum_{\beta \in \mathcal{I}_-} \tilde{c}_\beta S_\beta + \sum_{\gamma \in \mathcal{I}_+} \hat{c}_\gamma S_\gamma \right\} \\ &\quad - \left\{ \sum_{\beta \in \mathcal{I}_-} (c_\beta - \tilde{c}_\beta) S_\beta + \sum_{\gamma \in \mathcal{I}_+} (c_\gamma - \hat{c}_\gamma) S_\gamma \right\} . \end{aligned}$$

  By construction, the first term is still convex.

- To guarantee convexity, we have to bound the entropy contributions S_β in the second term for each $\beta \in \mathcal{I}_-$. To make this bound tighter, we include as many of the convex contributions S_γ as we can, while still satisfying the conditions in Theorem 4.2. Call the corresponding overcounting numbers $c_\gamma - \tilde{c}_\gamma \leq c_\gamma - \hat{c}_\gamma$ and put the remaining $\tilde{c}_\gamma - \hat{c}_\gamma$ back into the first term:

$$\begin{aligned} -S &= -\left\{ \sum_{\alpha} S_\alpha + \sum_{\beta \in \mathcal{I}_-} \tilde{c}_\beta S_\beta + \sum_{\gamma \in \mathcal{I}_+} \tilde{c}_\gamma S_\gamma \right\} \\ &\quad - \left\{ \sum_{\beta \in \mathcal{I}_-} (c_\beta - \tilde{c}_\beta) S_\beta + \sum_{\gamma \in \mathcal{I}_+} (c_\gamma - \tilde{c}_\gamma) S_\gamma \right\} . \end{aligned}$$

- Choose F_{convex} to be the first term plus a linear bound of the second term.

To find $\tilde{c}_\beta$ in the first step and similarly $\tilde{c}_\gamma$ in the third, we can use a linear program similar to the one described in Section 3.2 for checking the conditions of Theorem 3.1. We introduce slack variables θ_β and replace condition (7d) by

$$\sum_{\gamma \supset \beta} A_{\gamma\beta} = \theta_\beta \quad \forall_{\beta \in \mathcal{I}_-} \textit{(variable compensation)} ,$$

similar in spirit to (8). Furthermore, we add the inequality constraints $\theta_\beta \leq |c_\beta| \quad \forall_{\beta \in \mathcal{I}_-}$ (no need to compensate for more than $|c_\beta|$) and search for the maximum of $\theta \equiv \sum_{\beta \in \mathcal{I}_-} \theta_\beta$ (compensate as much as possible). In terms of the corresponding solution θ^*_β, we set $\tilde{c}_\beta = c_\beta - \theta^*_\beta$.

References

Aji, S., & McEliece, R. (2001). The generalized distributive law and free energy minimization. In *Proceedings of the Allerton Conference on Communication, Control, and Computing*.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society Series B*, *36*, 192–236.

Chiang, M., & Forney, G. (2001). Statistical physics, convex optimization and the sum product algorithm. Tech. rep., Stanford University.

Darroch, J., & Ratcliff, D. (1972). Generalized iterative scaling. *Annals of Mathematical Statistics*, *43*, 1470–1480.

Dechter, R., Kask, K., & Mateescu, R. (2002). Iterative join-graph propagation. In Darwiche, A., & Friedman, N. (Eds.), *Proceedings UAI-2002*, pp. 128–136.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, *39*, 1–38.

Hall, P. (1935). On representatives of subsets. *Journal of the London Mathematical Society*, *10*, 26–30.

Heckerman, D. (1989). A tractable inference algorithm for diagnosing multiple diseases. In Kanal, L., Henrion, M., Shachter, R., & Lemmer, J. (Eds.), *Proceedings of the Fifth Workshop on Uncertainty in Artificial Intelligence*, pp. 163–171, Amsterdam. Elsevier.

Heskes, T. (2003). Stable fixed points of loopy belief propagation are minima of the Bethe free energy. In Becker, S., Thrun, S., & Obermayer, K. (Eds.), *Advances in Neural Information Processing Systems 15*, pp. 359–366, Cambridge. MIT Press.

Heskes, T. (2004). On the uniqueness of loopy belief propagation fixed points. *Neural Computation*, *16*, 2379–2413.

Heskes, T., Albers, K., & Kappen, B. (2003). Approximate inference and constrained optimization. In *Uncertainty in Artificial Intelligence: Proceedings of the Nineteenth Conference (UAI-2003)*, pp. 313–320, San Francisco, CA. Morgan Kaufmann Publishers.

Heskes, T., Zoeter, O., & Wiegerinck, W. (2004). Approximate Expectation Maximization. In Thrun, S., Saul, L., & Schölkopf, B. (Eds.), *Advances in Neural Information Processing Systems 16*, pp. 353–360, Cambridge. MIT Press.

Ihler, A., Fisher, J., & Willsky, A. (2005). Loopy belief propagation: Convergence and effects of message errors. *Journal of Machine Learning Research*, *6*, 905–936.

Jaakkola, T., & Jordan, M. (1999). Variational probabilistic inference and the QMR-DT network. *Journal of Artificial Intelligence Research*, *10*, 291–299.

Jiroušek, R., & Přeučil, S. (1995). On the effective implementation of the iterative proportional fitting procedure. *Computational Statistics and Data Analysis*, *19*, 177–189.

Jordan, M., Ghahramani, Z., Jaakkola, T., & Saul, L. (1998). An introduction to variational methods for graphical models. In Jordan, M. (Ed.), *Learning in Graphical Models*, pp. 183–233. Kluwer Academic Publishers, Dordrecht.

Kikuchi, R. (1951). The theory of cooperative phenomena. *Physical Review*, *81*, 988–1003.

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. (1996). *Graphical models.* Oxford University Press, Oxford.

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 Gilliam, D., & Rosenthal, J. (Eds.), *Mathematical Systems Theory in Biology, Communications, Computation, and Finance*, pp. 275–300. Springer, New York.

Murphy, K. (2001). The Bayes Net toolbox for Matlab. *Computing Science and Statistics*, *33*, 331–350.

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

Neal, R., & Hinton, G. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In Jordan, M. (Ed.), *Learning in Graphical Models*, pp. 355–368. Kluwer Academic Publishers, Dordrecht.

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

Pakzad, P., & Anantharam, V. (2005). Estimation and marginalization using Kikuchi approximation methods. *Neural Computation*, *17*, 1836–1873.

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

Pelizzola, A. (2005). Cluster variation method in statistical physics and graphical models. *Journal of Physics A*, *38*, R309–R339.

Seung, S., Richardson, T., Lagarias, J., & Hopfield, J. (1998). Minimax and Hamiltonian dynamics of excitatory-inhibitory networks. In Jordan, M., Kearns, M., & Solla, S. (Eds.), *Advances in Neural Information Processing Systems 10*, pp. 329–335. MIT Press.

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

Teh, Y., & Welling, M. (2002). The unified propagation and scaling algorithm. In Dietterich, T., Becker, S., & Ghahramani, Z. (Eds.), *Advances in Neural Information Processing Systems 14*, pp. 953–960, Cambridge. MIT Press.

Wainwright, M., Jaakkola, T., & Willsky, A. (2002a). A new class of upper bounds on the log partition function. In Darwiche, A., & Friedman, N. (Eds.), *Uncertainty in Artificial Intelligence: Proceedings of the Eighteenth Conference (UAI-2002)*, pp. 536–543, San Francisco, CA. Morgan Kaufmann Publishers.

Wainwright, M., Jaakkola, T., & Willsky, A. (2002b). Tree-based reparameterization for approximate estimation on loopy graphs. In Dietterich, T., Becker, S., & Ghahramani, Z. (Eds.), *Advances in Neural Information Processing Systems 14*, pp. 1001–1008, Cambridge. MIT Press.

Wainwright, M., Jaakkola, T., & Willsky, A. (2003). Tree-reweighted belief propagation algorithms and approximate ML estimation via pseudo-moment matching. In Bishop, C., & Frey, B. (Eds.), *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*. Society for Artificial Intelligence and Statistics.

Yedidia, J., Freeman, W., & Weiss, Y. (2001). Generalized belief propagation. In Leen, T., Dietterich, T., & Tresp, V. (Eds.), *Advances in Neural Information Processing Systems 13*, pp. 689–695, Cambridge. MIT Press.

Yedidia, J., Freeman, W., & Weiss, Y. (2005). Constructing free energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, *51*, 2282–2312.

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