CENTER FOR
MACHINE PERCEPTION

CZECH TECHNICAL
UNIVERSITY IN PRAGUE

Zero-Temperature Limit of a Convergent Algorithm to Minimize the Bethe Free Energy

Tomáš Werner

CTU–CMP–2011–14

December 2011

Available at
ftp://cmp.felk.cvut.cz/pub/cmp/articles/werner/Werner-TR-2011-14.pdf

The work has been supported by the European Commission project FP7-ICT-270138 and the Czech Grant Agency project P103/10/0783.

Research Reports of CMP, Czech Technical University in Prague, No. 14, 2011

Published by

Center for Machine Perception, Department of Cybernetics
Faculty of Electrical Engineering, Czech Technical University
Technická 2, 166 27 Prague 6, Czech Republic
fax +420 2 2435 7385, phone +420 2 2435 7637, www: http://cmp.felk.cvut.cz

RESEARCH REPORT

ISSN 1213-2365

Zero-Temperature Limit of a Convergent Algorithm to Minimize the Bethe Free Energy

Tomáš Werner

December 2011

Abstract

After the discovery that fixed points of loopy belief propagation coincide with stationary points of the Bethe free energy, several researchers proposed provably convergent algorithms to directly minimize the Bethe free energy. These algorithms were formulated only for non-zero temperature (thus finding fixed points of the sum-product algorithm) and their possible extension to zero temperature is not obvious. We present the zero-temperature limit of the double-loop algorithm by Heskes, which converges a max-product fixed point. The inner loop of this algorithm is max-sum diffusion. Under certain conditions, the algorithm combines the complementary advantages of the max-product belief propagation and max-sum diffusion (LP relaxation): it yields good approximation of both ground states and max-marginals.

1 Introduction

Loopy belief propagation [17] is a well-known algorithm to approximate marginals of the Gibbs distribution defined by an undirected graphical model. For acyclic graphs, BP always converges and yields the exact marginals. For graphs with cycles, it is not guaranteed to converge but when it does, it often yields surprisingly good approximations of the true marginals. One informal argument for this is that at a BP fixed point, marginals are exact in every sub-tree of the factor graph [23, 24]. Attempts to understand loopy BP has generated a large body of literature, see e.g. the survey [25].

BP has a modification, known as the max-product BP, where summations are replaced with maximizations. In statistical mechanics terminology, this can be understood as the zero-temperature limit of the ordinary BP. Max-product BP computes (or approximates) max-marginals rather than ordinary marginals.

After the discovery [34, 33] that BP fixed points coincide with stationary points of the Bethe free energy, several researchers proposed provably convergent algorithms to find a local minimum of the Bethe free energy [35, 28, 22, 5, 6]. These algorithms have been proposed only for the sum-product and their possible extension to the max-product is not obvious.

We reformulate the double-loop algorithm [5] by Heskes such that taking its zero-temperature limit becomes straightforward, which results in an algorithm that always converges to a max-product BP fixed point. The inner loop of the algorithm is max-sum diffusion [13, 29, 31, 2]. We empirically observed that with a uniform initialization, the algorithm always yielded the same approximation of ground states that would be obtained by max-sum diffusion (or other algorithms for MAP inference based on LP relaxation, such as TRW-S [12]). Thus, it combines the complementary advantages of max-sum belief propagation and LP relaxation: unlike the former, it yields good approximation of ground states and, unlike the latter, it yields a good approximation of max-marginals.

The text is organized as follows. We first (§2) review the basics of inference in graphical models. We thoroughly discuss the zero-temperature limit of the Gibbs distribution and related quantities and how to obtain their approximation by variational inference. Then we review two basic cases of variational inference, with a convex free energy (§3) and with the Bethe free energy (§4). Then (§3.1, §4.1) we discuss their zero-temperature limits in detail. Finally (§5) we reformulate the double-loop algorithm [5] and modify it for the zero temperature.

2 Gibbs distribution

Let V be a set of variables, each variable $v \in V$ taking states x_v from a finite domain X_v. An assignment to a variable subset $a \subseteq V$ is $x_a \in X_a$, where X_a is the Cartesian product of domains X_v for $v \in a$. In particular, $x_V \in X_V$ is an assignment to all the variables. Let $E \subseteq 2^V$, thus (V, E) is a hypergraph. Each variable $v \in V$ and hyperedge $a \in E$ is assigned a potential function $\theta_v \colon X_v \to \overline{\mathbb{R}}$ and $\theta_a \colon X_a \to \overline{\mathbb{R}}$, respectively, where $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty\}$. All numbers $\theta_v(x_v)$ and $\theta_a(x_a)$ are understood as a single vector $\theta \in \overline{\mathbb{R}}^I$ (or mapping $\theta \colon I \to \overline{\mathbb{R}}$) with

$$I = \{\, (v, x_v) \mid v \in V,\ x_v \in X_v \,\} \cup \{\, (a, x_a) \mid a \in E,\ x_a \in X_a \,\}.$$

The Gibbs probability distribution over the hypergraph (V, E) is given by

$$p(x_V) = \exp[\, \langle \theta, \delta(x_V) \rangle - \Phi(\theta) \,] \tag{1}$$

where the mapping $\delta \colon X_V \to \{0, 1\}^I$ is such that

$$\langle \theta, \delta(x_V) \rangle = \sum_{v \in V} \theta_v(x_v) + \sum_{a \in E} \theta_a(x_a). \tag{2}$$

For infinite weights, we set $-\infty \cdot 0 = 0$ in the scalar product $\langle \theta, \delta(x_V) \rangle$. Since unary terms are included in (2) explicitly, we assume that E contains no singletons. The distribution is normalized by the *log-partition function*

$$\Phi(\theta) = \log \sum_{x_V} \exp \langle \theta, \delta(x_V) \rangle = \bigoplus_{x_V} \langle \theta, \delta(x_V) \rangle. \tag{3}$$

In (3), we used $x \oplus y = \log(\mathrm{e}^x + \mathrm{e}^y)$ to denote the *log-sum-exp operation*. It will be useful to keep in mind algebraic properties of this operation. It is associative and commutative, and addition distributes over it. Thus, $(\overline{\mathbb{R}}, \oplus, +)$ is a commutative semiring. This semiring is, via the logarithm map, isomorphic to the 'sum-product' semiring $(\mathbb{R}_+, +, \times)$.

Marginals. The marginals of the distribution are

$$\mu_v(x_v) = \sum_{x_{V\setminus v}} p(x_V), \quad \mu_a(x_a) = \sum_{x_{V\setminus a}} p(x_V), \tag{4}$$

where we abuse notation by writing $V \setminus v$ instead of $V \setminus \{v\}$. The numbers (4) are understood as a vector $\mu \in [0,1]^I$. All realizable marginal vectors μ form the *marginal polytope* $\operatorname{conv} \delta(X_V)$, where $\delta(X_V) = \{\, \delta(x_V) \mid x_V \in X_V \,\}$. Besides (an exponential number of) other constraints, μ satisfies normalization and marginalization constraints

$$\sum_{x_v} \mu_v(x_v) = \sum_{x_a} \mu_a(x_a) = 1, \quad \sum_{x_{a\setminus v}} \mu_a(x_a) = \mu_v(x_v). \tag{5}$$

All vectors $\mu \geq 0$ satisfying (5) form the *local marginal polytope* Λ. We have $\operatorname{conv} \delta(X_V) \subseteq \Lambda$, with equality if and only if hypergraph (V, E) is acyclic (i.e., its factor graph is a tree). We also introduce a symbol for log-marginals,

$$\nu_a(x_a) = \log \mu_a(x_a) = \bigoplus_{x_{V\setminus a}} \langle \theta, \delta(x_V) \rangle - \Phi(\theta) \tag{6}$$

(and similarly for $\nu_v(x_v)$). For log-marginals, constraints (5) read

$$\bigoplus_{x_v} \nu_v(x_v) = \bigoplus_{x_a} \nu_a(x_a) = 0, \quad \bigoplus_{x_{a\setminus v}} \nu_a(x_a) = \nu_v(x_v). \tag{7}$$

Reparameterizations. A *reparameterization* is an affine transformation of vector θ that preserves (2) for all assignments $x_V \in X_V$. We first define the *local reparameterization* on a pair (a, v) as follows: subtract an arbitrary unary function $\alpha_{av}\colon X_v \to \mathbb{R}$ from θ_v and add the same function to θ_a,

$$\theta_v(x_v) \leftarrow \theta_v(x_v) - \alpha_{av}(x_v), \quad \theta_a(x_a) \leftarrow \theta_a(x_a) + \alpha_{av}(x_v). \tag{8}$$

This preserves (2) because α_{av} cancels out. We understand (8) as 'passing a message' α_{av}. Applying local reparameterization (8) to all pairs (a, v) with $v \in a \in E$ yields the general reparameterization

$$\theta^\alpha_v(x_v) = \theta_v(x_v) - \sum_{a \ni v} \alpha_{av}(x_v), \quad \theta^\alpha_a(x_a) = \theta_a(x_a) + \sum_{v \in a} \alpha_{av}(x_v) \tag{9}$$

where $\alpha = \{\, \alpha_{av}(x_v) \mid v \in a \in E,\ x_v \in X_v \,\}$ is the vector of all messages and the transformed vector θ is denoted $\theta^\alpha \in \overline{\mathbb{R}}^I$. Thus $\langle \theta^\alpha, \delta(x_V) \rangle = \langle \theta, \delta(x_V) \rangle$. In fact, we have more generally $\langle \theta^\alpha, \mu \rangle = \langle \theta, \mu \rangle$ for all μ satisfying (5) and all α.

Reparameterizations can be done either by directly modifying the vector θ or by keeping θ unchanged and storing the messages α. While the former may be better for theoretical analysis, the latter is preferable in algorithms. In the sequel we freely switch between these two views.

2.1 Zero-temperature limit

In this section, we will use $p(x_V \mid \theta)$ and $\mu_v(x_v \mid \theta)$, $\mu_a(x_a \mid \theta)$ to explicitly denote the dependence of distribution (1) and its marginals on θ.

In statistical physics, the Gibbs distribution is usually considered in a more general form as $p(x_V \mid \beta\theta)$, where $\beta > 0$ is the inverse temperature [16]. The limit $\beta \to \infty$ is then known as the *zero-temperature limit*.

It is elementary to show that the distribution

$$p^\infty(x_V \mid \theta) = \lim_{\beta\to\infty} p(x_V \mid \beta\theta) \tag{10}$$

is zero everywhere except at *ground states*, which are the maximizers of $p(x_V \mid \theta)$ or, equivalently, $\langle \theta, \delta(x_V)\rangle$. If there are multiple ground states then the mass is distributed evenly among them.

The zero-temperature limit of the log-partition function is

$$\Phi^\infty(\theta) = \lim_{\beta\to\infty} \frac{\Phi(\beta\theta)}{\beta} = \max_{x_V} \langle \theta, \delta(x_V)\rangle, \tag{11}$$

which follows from the limit

$$\lim_{\beta\to\infty} \frac{(\beta x) \oplus (\beta y)}{\beta} = \max\{x, y\}. \tag{12}$$

The zero-temperature limit of log-marginals (6) yields *max-marginals*[1]

$$\nu_a^\infty(x_a) = \lim_{\beta\to\infty} \frac{\nu_a(x_a \mid \beta\theta)}{\beta} = \max_{x_{V\setminus a}} \langle \theta, \delta(x_V)\rangle - \Phi^\infty(\theta) \tag{13}$$

(similarly for $\nu_v^\infty(x_v)$). Observe that (13) and (11) differs from (6) and (3) only by replacing the log-sum-exp operation '$\oplus$' with 'max'. This corresponds, by the limit (12), to transition from the semiring $(\overline{\mathbb{R}}, \oplus, +)$ to the max-sum semiring $(\overline{\mathbb{R}}, \max, +)$. Similarly, max-marginals satisfy normalization and marginalization conditions (7) in which '$\oplus$' has been replaced with 'max'.

Max-marginals should not be confused[2] with the marginals of $p^\infty(x_V \mid \theta)$. These are different quantities and one cannot be computed from the other.

[1] It would be more precise to call (13) 'max-log-marginals' or 'log-max-marginals'. But, as is usual in the literature, we call them only 'max-marginals'.

[2] Unlike the limit (10), the limit (13) from (log-)marginals to max-marginals rarely appears in the machine learning or computer vision literature. The only work we found is [8].

Recovering ground states from max-marginals. Ground states can be recovered from max-marginals. To show that, we first recall what is the *constraint satisfaction problem* (CSP) [15, 1]. The CSP instance is defined by a vector $c \in \{0,1\}^I$, where functions $c_v\colon X_v \to \{0,1\}$ and $c_a\colon X_a \to \{0,1\}$ are understood as relations. A solution of the CSP is an assignment x_V satisfying all the relations, i.e., $c_v(x_v) = 1$ for all $v \in V$ and $c_a(x_a) = 1$ for all $a \in E$.

For a vector $\theta \in \overline{\mathbb{R}}^I$ we define vector $\lceil\theta\rceil \in \{0,1\}^I$ by

$$\lceil\theta\rceil_v(x_v) = \begin{cases} 1 & \text{if } x_v \in \operatorname*{argmax}_{y_v} \theta_v(y_v) \\ 0 & \text{otherwise} \end{cases}, \quad \lceil\theta\rceil_a(x_a) = \begin{cases} 1 & \text{if } x_a \in \operatorname*{argmax}_{y_a} \theta_a(y_a) \\ 0 & \text{otherwise} \end{cases},$$

i.e., a component of $\lceil\theta\rceil$ equals 1 iff the corresponding component of θ is maximal in its potential function. We say that such a components of θ is *active*. Now the set $\operatorname{argmax}_{x_V} \langle\theta, \delta(x_V)\rangle$ of ground states is the solution set of the CSP defined by vector $\lceil\nu^\infty\rceil$ of active max-marginals.

2.2 Convex conjugacy and variational inference

Let $H(\mu)$ denote the entropy of the distribution (1) as a function of its marginals. The functions $\Phi(\theta)$ and $-H(\mu)$ are convex and they are related by convex conjugacy,

$$\Phi(\theta) = \max_{\mu \in \operatorname{conv}\delta(X_V)} [\langle\theta, \mu\rangle + H(\mu)], \tag{14}$$

where the optimum is attained for μ equal to the marginals (4). In statistical physics, the quantity $-\langle\theta,\mu\rangle - H(\mu)$ is known as the *Gibbs free energy* of the system. By taking the limit $\beta \to \infty$ of the expression

$$\frac{\Phi(\beta\theta)}{\beta} = \max_{\mu \in \operatorname{conv}\delta(X_V)} \left[\langle\theta, \mu\rangle + \frac{H(\mu)}{\beta}\right] \tag{15}$$

we similarly obtain $\Phi^\infty(\theta)$ and max-marginals ν^∞.

The idea behind *variational inference* [25] is to replace the marginal polytope $\operatorname{conv}\delta(X_V)$ and the entropy $H(\mu)$ in (14) with their tractable approximations. Then the optimal value and the optimal argument of (14) is an approximation of the log-partition function and marginals, respectively. For $\beta \to \infty$,

- the optimal value of (15) is an approximation of $\Phi^\infty(\theta)$,
- the logarithm of the optimal argument of (15) is an approximation of max-marginals ν^∞,
- the solution set of the CSP defined by active approximate max-marginals is an approximation of the set $\operatorname{argmax}_{x_V} \langle\theta, \delta(x_V)\rangle$ of ground states.

As the entropy term in (15) approaches 0 for $\beta \to \infty$, one may think that it could be simply omitted. However, as pointed out in [27], if the approximate entropy is non-convex (such as the Bethe entropy), the problem (15) can have multiple local minima for arbitrarily large β. Thus, if our algorithm finds only a local minimum of (15), the entropy term is crucial.

3 Convex free energy

Let the marginal polytope in (14) be approximated by the local marginal polytope Λ and the true entropy by $H(\mu) \approx -\langle \log \mu, \mu \rangle$. This entropy approximation is concave, thus we obtained a simple (arguably, the simplest possible) variational inference method with a convex free energy [27, 4]. The approximation of (14) now reads

$$\max_{\mu \in \Lambda} \langle \theta - \log \mu, \mu \rangle. \tag{16}$$

The problem (16) can be solved as described e.g. in [32]. Its dual reads as follows: find a reparameterization of the original vector θ that minimizes the function

$$U(\theta) = \sum_{v \in V} \bigoplus_{x_v} \theta_v(x_v) + \sum_{a \in E} \bigoplus_{x_a} \theta_a(x_a). \tag{17}$$

This is a majorant of the log-partition function, $U(\theta) \geq \Phi(\theta)$ for every θ. A sufficient condition for dual optimality is that

$$\bigoplus_{x_{a \setminus v}} \theta_a(x_a) = \theta_v(x_v) \tag{18}$$

for all $v \in a \in E$ and $x_v \in X_v$. The primal and dual optimum are related by

$$\log \mu_v(x_v) = \theta_v(x_v) - \bigoplus_{y_v} \theta_v(y_v), \quad \log \mu_a(x_a) = \theta_a(x_a) - \bigoplus_{y_a} \theta_a(y_a). \tag{19}$$

Since function (17) is convex and differentiable, its global minimum over reparameterizations of θ can be found by coordinate descent. This leads to a simple message passing algorithm. The iteration of this algorithm enforces equality (18) for a single pair (a, v) by local reparameterization (8), which determines $\alpha_{av}(x_v)$ in (8) uniquely. The iteration decreases $U(\theta)$, and this decrease is strict unless $U(\theta)$ is already minimal. On convergence, (18) holds globally.

If reparameterizations are represented by messages rather than by directly modifying θ, the dual of (16) reads $\min_\alpha U(\theta^\alpha)$ and the coordinate descent becomes Algorithm 1. To correctly handle infinite weights, the algorithm expects that $[\theta_v(x_v) > -\infty] \Leftrightarrow [\max_{x_{a \setminus v}} \theta_a(x_a) > -\infty]$ for all $v \in a \in E$ and $x_v \in X_v$.

Algorithm 1 The ‘diffusion’ algorithm.

```
repeat
    for $v \in a \in E$ and $x_v \in X_v$ such that $\theta_v(x_v) > -\infty$ do
        $\alpha_{av}(x_v) \leftarrow \alpha_{av}(x_v) + \frac{1}{2}\Big[\theta_v^\alpha(x_v) - \bigoplus_{x_{a \setminus v}} \theta_a^\alpha(x_a)\Big]$
    end for
until convergence
```

3.1 Zero-temperature limit: max-sum diffusion

The zero-temperature limit of the optimization problem above is obtained by replacing θ with $\beta\theta$ and taking the limit $\beta \to \infty$. This results in replacing '$\oplus$' with 'max' in (17)–(19) and Algorithm 1. We assume that this has been done.

This yields the LP relaxation approach to maximizing the Gibbs distribution first proposed by Schlesinger et al. [19, 13], see also [29, 32, 31, 2]. In these works, the zero-temperature limit of Algorithm 1 is called *max-sum diffusion*.

Let function (17) after replacing '$\oplus$' with 'max' be denoted by

$$U^\infty(\theta) = \lim_{\beta\to\infty} \frac{U(\beta\theta)}{\beta} = \sum_{v\in V} \max_{x_v} \theta_v(x_v) + \sum_{a\in E} \max_{x_a} \theta_a(x_a).$$

We have $U^\infty(\theta) \geq \Phi^\infty(\theta)$ for every θ. Algorithm 1 tries to minimize $U^\infty(\theta)$ by reparameterizing θ. However, the function U^∞ is non-differentiable now – therefore Algorithm 1 may converge only to a local (with respect to coordinate moves) minimum of $U^\infty(\theta)$. While it is easy to prove convergence of the algorithm in value, convergence in argument is only a conjecture to date [29, 31] and only a weaker property has been proved recently [18].

According to §2.2, when θ is optimal then $U^\infty(\theta)$ is an approximation of $\Phi^\infty(\theta)$ and (19) is an approximation of the max-marginals ν^∞. Note that the approximate max-marginals (19) are, up to normalization, directly equal to θ. Since vector $\lceil\theta\rceil$ is not affected by normalization of θ, the solution set of the CSP $\lceil\theta\rceil$ is an approximation of the ground states.

But this is in agreement with [19, 29], where it is shown that the inequality $U^\infty(\theta) \geq \Phi^\infty(\theta)$ (and hence the LP relaxation) is tight if and only if the CSP defined by $\lceil\theta\rceil$ has a solution. Then, $\langle\theta, \delta(x_V)\rangle = \Phi^\infty(\theta)$ for every solution x_V of CSP $\lceil\theta\rceil$. There are two important problem subclasses for which the LP relaxation is tight: if hypergraph (V, E) is acyclic or if the functions θ_a are (permuted) supermodular [29, 31, 11]. Besides, it is tight for many other instances met in applications. This makes this method very suitable for approximating ground states, which has been also observed empirically[3] [21].

However, even when the LP relaxation is tight, (19) are a very poor approximation of max-marginals. They are inexact even for acyclic hypergraphs.

4 Bethe free energy and belief propagation

Let the true entropy in (14) be approximated by the *Bethe entropy*

$$H(\mu) \approx -\langle \log\mu, \mu\rangle + \sum_{v\in V} n_v \langle \log\mu_v, \mu_v\rangle, \tag{20}$$

where n_v is the number of hyperedges containing variable v. For acyclic hypergraphs the Bethe entropy is equal to $H(\mu)$, otherwise it can be non-concave and

[3] The TRW-S algorithm [12] studied in [21] solves the same LP relaxation as max-sum diffusion. The same holds for zero-temperature versions of other recently proposed convergent algorithms to minimize convex free energies [9, 3, 27, 4].

even negative on Λ. Then (14) reads

$$\max_{\mu\in\Lambda}\Big[\langle\theta-\log\mu,\mu\rangle+\sum_{v\in V}n_v\langle\log\mu_v,\mu_v\rangle\Big]. \tag{21}$$

The negative objective of (21) is the *Bethe free energy*.

Next we formulate loopy belief propagation. Unlike in the 'traditional' formulation [17, 14, 33, 25], we identify messages with reparameterizations, which agrees with [12, eq. (2)] and [30]. Let the marginals (4) be approximated as

$$\log\mu_v(x_v)=\hat\theta_v(x_v)-\bigoplus_{y_v}\hat\theta_v(y_v),\quad \log\mu_a(x_a)=\hat\theta_a(x_a)-\bigoplus_{y_a}\hat\theta_a(y_a) \tag{22}$$

where

$$\hat\theta_v=\theta_v,\quad \hat\theta_a=\theta_a+\sum_{v\in a}\theta_v. \tag{23}$$

Note that μ_v and μ_a is the Gibbs distribution for the simple graphical model with hypergraph $(\{v\},\emptyset)$ and $(a,\{a\})$, respectively. This corresponds to decomposing (V,E) into small sub-hypergraphs. In general, μ fails to satisfy the local marginalization conditions of (5). Plugging (22) into these conditions yields

$$\bigoplus_{x_{a\setminus v}}\Big[\theta_a(x_a)+\sum_{u\in a}\theta_u(x_u)\Big]=\theta_v(x_v)+\mathrm{const}_{av}, \tag{24}$$

which by cancelling $\theta_v(x_v)$ simplifies to

$$\bigoplus_{x_{a\setminus v}}\Big[\theta_a(x_a)+\sum_{u\in a\setminus v}\theta_u(x_u)\Big]=\mathrm{const}_{av}. \tag{25}$$

Here, const_{av} is a constant independent on x_v. We define a *belief propagation fixed point* to be a vector θ satisfying (25) for all $v\in a\in E$ and $x_v\in X_v$. The BP algorithm then tries to reparameterize θ to make it satisfy (25).

As discovered by Yedidia et al. [33], BP fixed points (25) correspond to stationary points of problem (21) via the map (22). Heskes [5] showed that every *stable* BP fixed point is a local maximum (rather than minimum or saddle point) of (21), but not necessarily *vice versa.*

4.1 Zero-temperature limit: max-sum belief propagation

In the zero-temperature limit, '$\oplus$' in (22)–(25) is replaced with 'max'. We assume in §4.1 that this has been done. Then, (25) defines a fixed point of *max-sum belief propagation*[4].

[4] The traditional names 'sum-product' and 'max-product' are misnomers in our paper because we stated BP in the logsumexp-sum semiring $(\overline{\mathbb{R}},\oplus,+)$ rather than (as is usual) in the (isomorphic) sum-product semiring $(\mathbb{R}_+,+,\times)$. For zero temperature, we are then in the max-sum semiring $(\overline{\mathbb{R}},\max,+)$ rather than in the max-product semiring $(\mathbb{R}_+,\max,\times)$.

According to §2.2, numbers (22) are approximations of max-marginals ν^∞ and the solution set of the CSP defined by active approximate max-marginals is an approximation of the set $\operatorname{argmax}_{x_V} \langle\theta, \delta(x_V)\rangle$ of ground states[5]. Since approximate max-marginals (22) are, up to normalization, equal to numbers (23), this CSP is defined by $\lceil\hat{\theta}\rceil$. This formulation is consistent because (as is easy to verify) the value $\langle\theta, \delta(x_V)\rangle$ is the same for all solutions x_V of the CSP $\lceil\hat{\theta}\rceil$.

It is well-known that the approximation of ground states obtained by max-sum belief propagation is often poor (letting alone that the algorithm may not converge). In our formalism, the value $\langle\theta, \delta(x_V)\rangle$ for the solutions x_V of CSP $\lceil\hat{\theta}\rceil$ are often far[6] from $\Phi^\infty(\theta)$. It may of course also happen that the CSP $\lceil\hat{\theta}\rceil$ has no solution. The situation is especially intriguing if the functions θ_a are supermodular[7]. Then maximizing $\langle\theta, \delta(x_V)\rangle$ is tractable but the approximation obtained from max-sum BP can be inexact [25].

On the other hand, if the approximation of ground states from max-sum BP is good, then usually also the approximation (22) of max-marginals is good. This is intuitively justified by the fact that at a BP fixed point, the (max-)marginals are exact in every subtree of the factor graph [23, 24].

5 Direct minimization of the Bethe free energy

Heskes [5, 6] proposed a class of convergent algorithms to find a local minimum of Bethe and Kikuchi free energies, based on the *minorize-maximize approach* [7, 20]. We now describe a simple representant of this class, which finds a local maximum of the non-concave maximization problem (21).

Let $F(\mu)$ denote the objective of (21). A family of minorants of F is constructed as

$$\tilde{F}(\mu, \tilde{\mu}) = \langle\theta - \log\mu, \mu\rangle + \sum_{v\in V} n_v \langle\log\tilde{\mu}_v, \mu_v\rangle, \tag{26}$$

where $\tilde{\mu}$ is a collections of variable distributions $\tilde{\mu}_v$, non-negative and normalized. For any μ and $\tilde{\mu}$ we have $\tilde{F}(\mu, \tilde{\mu}) \le F(\mu)$, with equality if and only if $\tilde{\mu}_v = \mu_v$ for all $v \in V$. This follows from the well-known fact that any non-negative and normalized vectors μ_v and $\tilde{\mu}_v$ satisfy $\langle\log\tilde{\mu}_v, \mu_v\rangle \le \langle\log\mu_v, \mu_v\rangle$, which holds with equality only if $\tilde{\mu}_v = \mu_v$.

The problem (21) is now split into two nested problems

$$\max_{\tilde{\mu}} \max_{\mu\in\Lambda} \tilde{F}(\mu, \tilde{\mu}). \tag{27}$$

[5] Decoding an assignment from a fixed point of the loopy max-sum/max-product BP has been addressed in the BP literature (see e.g. [26, 25]) but never has been formulated as a CSP. But we believe this is a very natural formulation.

[6] Note that this never happens for max-sum diffusion, where solutions of the CSP $\lceil\theta\rceil$ are inevitably ground states.

[7] For supermodular θ_a, CSP $\lceil\hat{\theta}\rceil$ always has a solution. This is easy to prove: since function θ_a are supermodular, functions $\hat{\theta}_a$ are supermodular as well, and then the proof proceeds like the proof [31] that max-sum diffusion exactly solves (permuted) supermodular problems.

The inner problem is a concave maximization, which can be solved optimally – in fact, it has the form (16). The objective $\max_{\mu\in\Lambda} \tilde{F}(\mu,\tilde{\mu})$ of the outer problem is a non-concave function of $\tilde{\mu}$ and thus we can only hope to find its local maximum. The algorithm has two nested loops, corresponding to the inner and outer problem. The outer iteration has two steps:

1. Keeping $\tilde{\mu}$ fixed, find $\mu \in \Lambda$ that maximizes $\tilde{F}(\mu,\tilde{\mu})$.
2. For all $v \in V$, set $\tilde{\mu}_v \leftarrow \mu_v$.

Each of these two steps increases $\tilde{F}(\mu,\tilde{\mu})$. For Step 1, this is true by definition. For Step 2, it follows from the minorization property of $\tilde{F}$. The algorithm converges to a state when μ is the maximum of $\tilde{F}(\mu,\tilde{\mu})$ and $\tilde{\mu}_v = \mu_v$, therefore μ is a local maximum of (21).

In Step 1, $\tilde{F}(\mu,\tilde{\mu})$ needs to be maximized over $\mu \in \Lambda$. This can be cast in the form (16). First we substitute $\log\tilde{\mu} = \tilde{\theta}$. Note that after this substitution, the normalization condition $\sum_{x_v} \tilde{\mu}_v(x_v) = 1$ reads $\bigoplus_{x_v} \tilde{\theta}_v(x_v) = 0$. Then

$$\tilde{F}(\mu,\tilde{\mu}) = \langle \theta - \log\mu, \mu\rangle + \sum_{v\in V} n_v \langle \tilde{\theta}_v, \mu_v\rangle = \langle \hat{\theta} - \log\mu, \mu\rangle \tag{28}$$

where, using that $\sum_v n_v \tilde{\theta}_v = \sum_a \sum_{v\in a} \tilde{\theta}_v$, the vector $\hat{\theta}$ is given by[8]

$$\hat{\theta}_v = \theta_v, \quad \hat{\theta}_a = \theta_a + \sum_{v\in a} \tilde{\theta}_v. \tag{29}$$

The inner problem is dualized, which changes (27) to a saddle-point problem. As described in §3, the dual is solved by reparameterizing $\hat{\theta}$ such that $\hat{\theta}$ satisfies (18) (which minimizes $U(\hat{\theta})$) and then computing μ from $\hat{\theta}$ using (19). Since $\hat{\theta}^\alpha_a = \theta^\alpha_a + \sum_{v\in a} \tilde{\theta}_v$, we can reparameterize θ instead of $\hat{\theta}$. The outer iteration now reads as follows:

1. Reparameterize θ such that
$$\bigoplus_{x_{a\setminus v}} \Big[\theta_a(x_a) + \sum_{u\in a} \tilde{\theta}_u(x_u)\Big] = \theta_v(x_v). \tag{30}$$
2. For all $v \in V$, set $\tilde{\theta}_v \leftarrow \theta_v - \bigoplus\limits_{x_v} \theta_v(x_v)$.

The number

$$U(\hat{\theta}) = \sum_{v\in V} \bigoplus_{x_v} \theta_v(x_v) + \sum_{a\in E} \bigoplus_{x_a} \Big[\theta_a(x_a) + \sum_{v\in a} \tilde{\theta}_v(x_v)\Big]$$

is decreased by Step 1 and it is increased by Steps 1+2 combined. The algorithm converges to a state when $\tilde{\theta}_v = \theta_v - \bigoplus_{x_v} \theta_v(x_v)$. Then, θ is a BP fixed point. This is indeed very obvious: since $\tilde{\theta}_v$ and θ_v are equal up to an additive constant,

[8] We could alternatively choose $\hat{\theta}$ as $\hat{\theta}_v = \theta_v + n_v\tilde{\theta}_v$, $\hat{\theta}_a = \theta_a$. But since (29) directly compares to (23), the choice (29) more clearly shows the connection with BP fixed points.

(30) becomes the same as (24), therefore (25) holds. If reparameterizations are represented by messages, we obtain Algorithm 2.

Let us remark that the normalization in Step 2 is not necessary, we could just set $\tilde\theta_v \leftarrow \theta_v$. This would not affect convergence to a BP fixed point but $U(\hat\theta)$ would lose its meaning and $\tilde\theta_v$ might grow unbounded.

Algorithm 2 Double-loop algorithm to find a BP fixed point.

```
Initialization: Choose any $\tilde\theta$ with $\bigoplus_{x_v} \tilde\theta_v(x_v) = 0$. Choose any $\alpha$.
repeat                                                              ▷ outer loop
    repeat                                                          ▷ inner loop
        for $v \in a \in E$ and $x_v \in X_v$ such that $\theta_v(x_v) > -\infty$ do
            $\alpha_{av}(x_v) \leftarrow \alpha_{av}(x_v) + \frac{1}{2}\Big[\theta_v^\alpha(x_v) - \bigoplus_{x_{a\setminus v}} \Big[\theta_a^\alpha(x_a) + \sum_{u\in a} \tilde\theta_u(x_u)\Big]\Big]$
        end for
    until convergence
    For all $v \in V$, set $\tilde\theta_v \leftarrow \theta_v^\alpha - \bigoplus_{x_v} \theta_v^\alpha(x_v)$.
until convergence
```

The outer loop is guaranteed to converge only if the inner loop reaches full convergence. There is no theoretical guarantee ensuring convergence with a finite number of inner iterations – this unpleasant feature is common to double-loop algorithms applied to saddle-point problems. However, this does not seem to be an issue in practice.

5.1 Zero-temperature limit

Replacing θ with $\beta\theta$ in all the formulas and taking the limit $\beta \to \infty$ again results in replacing '$\oplus$' with 'max'. Then, Algorithm 2 converges to a max-sum belief propagation fixed point.

Though we never observed the algorithm fail to converge, its convergence (with the inner loop run to full convergence) is only a conjecture. The argument is that if it converges for any $\beta < \infty$ then it is reasonable to assume that it will converge also in the limit. But we suspect that finding a formal proof for $\beta \to \infty$ may be difficult, especially when convergence of the inner loop (max-sum diffusion) itself is a conjecture to date. Note that, unlike for $\beta < \infty$, the proof cannot be based on the fact that the value of $U^\infty(\hat\theta)$ monotonically decreases because it often remains constant after the first several outer iterations.

Uniform initialization. Depending on the initial $\tilde\theta_v$, the algorithm can converge to different fixed points (as we indeed observed). Particularly interesting is the case when the initial $\tilde\theta_v$ are all uniform – due to the normalization condition $\max_{x_v} \tilde\theta_v(x_v) = 0$, this means $\tilde\theta = 0$. Next we focus only on this case.

Figure 1 shows how the algorithm converged for different types of pairwise interactions and different types of graph. Occasionally (e.g., for repulsive inter-

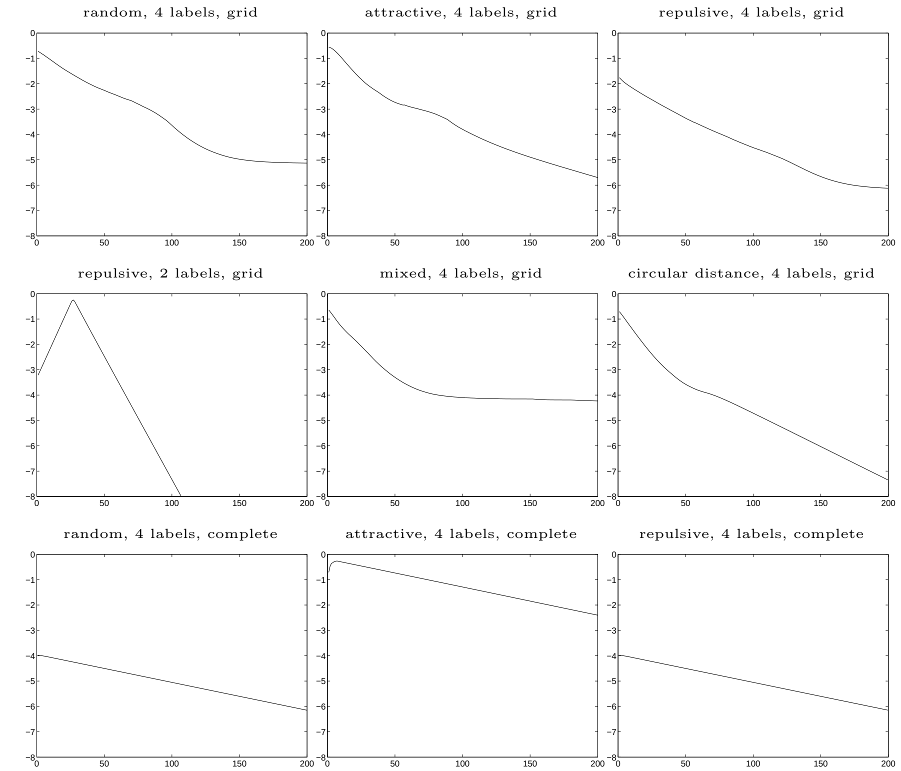

Figure 1: Convergence of zero-temperature version of Algorithm 2 with initial $\tilde{\theta}_v = 0$. The horizontal axis is the number of outer iterations, the vertical axis is $\log_{10}$ of average residuals to the max-sum BP fixed point condition (24). The title is in the form '*type of pairwise interactions, number of labels, graph type*'. The grid graph had 20×20 and the complete graph 40 vertices. The unary potentials were generated randomly.

actions and two labels) the residuals approached zero non-monotonically. The inner iteration was run to almost full convergence, however the results were not qualitatively affected by this.

We made the following key observation:

> *If the algorithm is initialized with $\tilde{\theta} = 0$ then after the first outer iteration $\lceil \hat{\theta} \rceil$ and $U^\infty(\hat{\theta})$ remain unchanged.*

This observation is only empirical, currently we have neither a formal proof nor a counterexample. It has an important consequence. If initially $\tilde{\theta} = 0$, then the first outer iteration is just Algorithm 1 applied to $\hat{\theta} = \theta$. If all subsequent outer iterations do not change $\lceil \hat{\theta} \rceil$, then CSP $\lceil \hat{\theta} \rceil$ after convergence of Algorithm 2 is the same as CSP $\lceil \theta \rceil$ that would be obtained by running Algorithm 1 on θ.

Thus, the approximate ground states obtained by Algorithm 2 are the same as those obtained by Algorithm 1. However, since Algorithm 2 converges to a max-sum BP fixed point, approximate max-marginals obtained by Algorithm 2 are expected to be much more accurate than those obtained by Algorithm 1.

6 Conclusion

We showed in §3.1 and §4.1 that the properties of max-sum diffusion (and all MAP inference algorithms based on LP relaxation) and max-sum belief propagation are complementary: the former yields good approximation of ground states but poor approximation of max-marginals, the latter *vice versa*. The double-loop algorithm initialized with $\tilde{\theta} = 0$ combines advantages of both: it yields approximate ground states that are exact for supermodular problems and approximate max-marginals that are exact in every sub-tree of the factor graph.

Our paper is primarily theoretical, more experiments are needed to compare approximate max-marginals from the double loop algorithm with ground truth.

We have not pursued another potentially interesting application of the double-loop algorithm with non-uniform initialization $\tilde{\theta} \neq 0$. It is known that max-sum BP occasionally yields better approximate ground states than LP relaxation. This has been observed e.g. for some problems on highly connected graphs [10]. However, the max-sum BP algorithm does not always converge, thus the convergent double loop algorithm might be useful here.

The double-loop algorithm could be speeded up by using an inner loop with tree updates as in e.g. TRW-S [12] rather than edge updates as in max-sum diffusion. We believe this is possible.

Acknowledgments

The work has been supported by the European Commission project FP7-ICT-270138 and the Czech Grant Agency project P103/10/0783. The author thanks Alexander Shekhovtsov for discussions.

References

[1] David Cohen and Peter Jeavons. The complexity of constraint languages. In *Handbook of Constraint Programming*, chapter 8. Elsevier, 2006.

[2] V. Franc, S. Sonnenburg, and T. Werner. Cutting plane methods in machine learning. In S. Sra, S. Nowozin, and S. J. Wright, editors, *Optimization for Machine Learning*. MIT Press, 2011.

[3] Amir Globerson and Tommi Jaakkola. Fixing max-product: Convergent message passing algorithms for MAP LP-relaxations. In *Neural Information Processing Systems (NIPS)*, pages 553–560, 2008.

[4] Tamir Hazan and Amnon Shashua. Convergent message-passing algorithms for inference over general graphs with convex free energies. In *Conf. on Uncertainty in Artificial Intelligence (UAI)*, pages 264–273, 2008.

[5] Tom Heskes. Stable fixed points of loopy belief propagation are minima of the Bethe free energy. In *Adv. in Neural Information Processing Systems (NIPS)*, pages 359–366, 2003.

[6] Tom Heskes. Convexity arguments for efficient minimization of the Bethe and Kikuchi free energies. *Jr. of Artificial Intelligence Research*, 26:153–190, 2006.

[7] D. R. Hunter and K. Lange. A tutorial on MM algorithms. *The American Statistician*, 58(1):30–38, 2004.

[8] Jason K. Johnson, Dmitry M. Malioutov, and Alan S. Willsky. Lagrangian relaxation methods for intractable graphical models. Presentation at Stochastic Systems Group (SSG) seminar at MIT, slides on `http://ssg.mit.edu/~jasonj/johnson-lr-ssg05.pdf`, September 2005.

[9] Jason K. Johnson, Dmitry M. Malioutov, and Alan S. Willsky. Lagrangian relaxation for MAP estimation in graphical models. In *Allerton Conf. Communication, Control and Computing*, 2007.

[10] V. Kolmogorov and C. Rother. Comparison of energy minimization algorithms for highly connected graphs. In *European Conf. Computer Vision (ECCV)*, pages II: 1–15, 2006.

[11] V. N. Kolmogorov and M. J. Wainwright. On the optimality of tree-reweighted max-product message-passing. In *Conf. Uncertainty in Artificial Intelligence (UAI)*, 2005.

[12] Vladimir Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 28(10):1568–1583, 2006.

[13] V. A. Kovalevsky and V. K. Koval. A diffusion algorithm for decreasing the energy of the max-sum labeling problem. Glushkov Institute of Cybernetics, Kiev, USSR. Unpublished, approx. 1975.

[14] F. R. Kschischang, B. J. Frey, and H. A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Trans. Information Theory*, 47(2):498–519, 2001.

[15] A. Mackworth. Constraint satisfaction. In *Encyclopaedia of Artificial Intelligence*, pages 285–292. Wiley, 1991.

[16] Marc Mezard and Andrea Montanari. *Information, Physics, and Computation.* Oxford University Press, USA, 2009.

[17] Judea Pearl. *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* Morgan Kaufmann, San Francisco, 1988.

[18] M. Schlesinger and K. Antoniuk. Diffusion algorithms and structural recognition optimization problems. *Cybernetics and Systems Analysis*, 47:175–192, 2011.

[19] M. I. Shlezinger. Syntactic analysis of two-dimensional visual signals in noisy conditions. *Cybernetics and Systems Analysis*, 12(4):612–628, 1976. Translation from Russian.

[20] Bharath Sriperumbudur and Gert Lanckriet. On the convergence of the concave-convex procedure. In *Adv. in Neural Information Processing Systems (NIPS)*, pages 1759–1767. 2009.

[21] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwal, M. Tappen, and C. Rother. A comparative study of energy minimization methods for Markov random fields. In *Eur. Conf. Computer Vision (ECCV)*, pages II: 16–29, 2006.

[22] Yee Whye Teh and Max Welling. The unified propagation and scaling algorithm. In *Conf. on Neural Information Processing Systems (NIPS)*, pages 953–960, 2001.

[23] M. Wainwright, T. Jaakkola, and A. Willsky. Tree-based reparameterization framework for analysis of sum-product and related algorithms. *IEEE Trans. Information Theory*, 49(5):1120–1146, 2003.

[24] M. Wainwright, T. Jaakkola, and A. Willsky. Tree consistency and bounds on the performance of the max-product algorithm and its generalizations. *Statistics and Computing*, 14:143–166, 2004.

[25] Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

[26] Yair Weiss and William T. Freeman. On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs. *IEEE Trans. Information Theory*, 47(2):736–744, 2001.

[27] Yair Weiss, Chen Yanover, and Talya Meltzer. MAP estimation, linear programming and belief propagation with convex free energies. In *Conf. on Uncertainty in Artificial Intelligence (UAI)*, 2007.

[28] Max Welling and Yee Whye Teh. Belief optimization for binary networks: A stable alternative to loopy belief propagation. In *Conf. on Uncertainty in Artificial Intelligence (UAI)*, pages 554–561, 2001.

[29] Tomáš Werner. A linear programming approach to max-sum problem: A review. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(7):1165–1179, July 2007.

[30] Tomáš Werner. Primal view on belief propagation. In *Conf. on Uncertainty in Artificial Intelligence (UAI)*, pages 651–657, July 2010.

[31] Tomáš Werner. Revisiting the linear programming relaxation approach to Gibbs energy minimization and weighted constraint satisfaction. *IEEE Trans. Pattern Analysis and Machine Intelligence*, August 2010.

[32] Tomáš Werner and Alexander Shekhovtsov. Unified framework for semiring-based arc consistency and relaxation labeling. In *12th Computer Vision Winter Workshop, St. Lambrecht, Austria*, pages 27–34. Graz University of Technology, February 2007.

[33] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Trans. Information Theory*, 51(7):2282–2312, 2005.

[34] Jonathan Yedidia, William T. Freeman, and Yair Weiss. Generalized belief propagation. In *Neural Information Processing Systems (NIPS)*, pages 689–695, 2000.

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