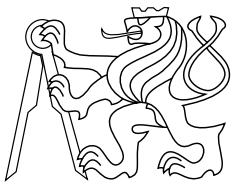
CENTER FOR
MACHINE PERCEPTION



CZECH TECHNICAL
UNIVERSITY IN PRAGUE

RESEARCH REPORT

ISSN 1213-2365

Fixed Points of Loopy Belief Propagation as Zero Gradients of a Function of Reparameterizations

Tomáš Werner

CTU–CMP–2010–05

February 2010

Available at
ftp://cmp.felk.cvut.cz/pub/cmp/articles/werner/Werner-TR-2010-05.pdf

This work was supported by the European Commission grant 215078
and the Czech government grant MSM6840770038.

Research Reports of CMP, Czech Technical University in Prague, No. 5, 2010

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

Fixed Points of Loopy Belief Propagation as Zero Gradients of a Function of Reparameterizations

Tomáš Werner

February 2010

Abstract

The existing view on loopy belief propagation sees it as an algorithm to find a common zero of a system of non-linear functions, not explicitly related to each other. We show that these functions are in fact related – they are the partial derivatives of a single function of reparameterizations. Thus, belief propagation searches for a zero gradient of a single function. We show that belief propagation fixed points are in one-to-one correspondence with zero gradient points of this function and that every such point is a saddle.

Contents

1 Introduction

Loopy belief propagation (further only belief propagation, BP) [7] is a well-known algorithm to approximate marginals and the partition function of the Gibbs probability distribution defined on an undirected graphical model (a Markov random field). For trees it yields the exact result, for graphs with cycles it often yields a surprisingly good approximation. A large literature exists on BP and related topics and we refer the reader to the recent survey by Wainwright and Jordan [12].

Unfortunately, BP on cyclic graphs is not guaranteed to converge, which is indeed often observed. A lot of effort has been invested to understanding this phenomenon (see [12, §4.1.3] for references). Solid ground was provided by the discovery by Yedidia et al. [18, 17] that BP fixed points coincide with the stationary points of the Bethe variational problem, long known in statistical physics. This problem minimizes a non-convex function (the Bethe free energy) of beliefs subject to the constraint that they satisfy the normalization and marginalization conditions. Heskes [3] showed that every stable BP fixed point is a local optimum (rather than a saddle point) of this problem, but not *vice versa*. Unfortunately, this did not entirely explain the BP algorithm itself because it does not directly solve the Bethe variational problem – although BP is an algorithm to solve the non-linear equation system describing the fixed points of the Bethe variational problem, it does not provide a feasible solution to this problem before it has converged.

The basic operation in the BP algorithm is 'passing a message', which means sending a vector of numbers between a node and an edge of the graph [7]. Messages turned out to be directly related to the Lagrange multipliers of the Bethe variational problem [18, 17]. Later it became clear [10] that passing a message corresponds to reparameterizing the distribution. In this view, BP tries to reparameterize the distribution so that the corresponding beliefs have consistent marginals.

Though this is generally known, no existing theoretical analysis of BP fully utilizes the interpretation of messages and the Lagrange multipliers of the Bethe variational problem as reparameterizations. As a minor contribution, we incorporate reparameterizations into variational inference and BP in a principled way, which makes the picture of BP clearer and more complete.

The current view on BP sees it as an algorithm to find a common zero of a set of functions, not explicitly related to each other. Our contribution is the observation that these functions are strongly related – they are the partial derivatives of a single function of reparameterizations. Thus, BP searches for a zero gradient of a single function. We show that BP fixed points are in one-to-one correspondence with zero gradient points of this function and that every such point is a saddle[1].

We follow the terminology and notation used by Wainwright et al. [12]. The text is organized as follows. In §2 we review the exponential families of probability distributions and the variational approach to approximative inference in graphical models. In §3 we revisit variational inference with a simple concave entropy approximation, derive the dual problem, the optimality conditions, and a coordinate descent algorithm to solve the dual – the finite-temperature version of max-sum diffusion [6, 15]. We include this section in order to contrast it in §4 with the non-concave Bethe entropy approximation. In §4, we recall the Bethe variational problem, show its dual has no explicit form, and prove the result [17] that its stationary points correspond to BP fixed points. In §5 we present our contribution, a function of reparameterization the zero gradients of which corresponds to BP fixed points, and give detailed properties of this function. We conclude in §6.

§2–§4 are rather detailed and therefore they have also a certain tutorial value.

[1] These saddles should not be confused with the saddle points in the convergent double-loop algorithms to minimize the Bethe free energy by Heskes [3, 4].

2 Exponential family of probability distributions

Let X and I be finite sets and $\phi\colon X \to \mathbb{R}^I$. The discrete natural exponential family is a family of probability distributions $p(\cdot\,|\,\theta)\colon X \to \mathbb{R}_{++}$ parameterized by canonical parameters[2] $\theta \in \mathbb{R}^I$, where

$$p(x\,|\,\theta) = \exp[\theta\phi(x) - F(\theta)] \tag{1}$$

Here, θ is a row vector and $\phi(x)$ a column vector, so that $\theta\phi(x) = \sum_{i\in I} \theta_i\phi_i(x)$. The component functions $\phi_i\colon X \to \mathbb{R}$ of ϕ are the basis functions of the family. The normalization term

$$F(\theta) = \bigoplus_{x\in X} \theta\phi(x) \tag{2}$$

is the convex *log-partition function* and $a \oplus b = \log(\mathrm{e}^a + \mathrm{e}^b)$ denotes the *log-sum-exp operation*.

Though it is not common, we find it convenient to use a special symbol for the log-sum-exp operation. Let us recall its key properties. It is the convex conjugate of entropy. It is associative and commutative and addition distributes over $\oplus$. It relates to addition the same way as addition relates to multiplication – more precisely, $(\oplus, +)$ is a commutative semiring over $\mathbb{R}$, isomorphic via the exp function to semiring $(+, \times)$ over $\mathbb{R}_{++}$.

2.1 Reparameterizations

If the functions ϕ_i are affinely independent they form a *minimal representation* of the family, otherwise they form an *overcomplete representation*. In the latter case, all affine dependencies among ϕ_i can be written as

$$A\phi(x) = 0, \quad B\phi(x) = 1 \qquad \forall x \in X \tag{3}$$

for some matrices A and B, where 0 and 1 are column vectors of zeros and ones. Matrix A captures homogeneous and matrix B inhomogeneous dependencies. If α and β are arbitrary row vectors and

$$\theta' = \theta + \alpha A + \beta B \tag{4}$$

then $\theta'\phi(x) = \theta\phi(x) + \beta 1$ and $F(\theta') = F(\theta) + \beta 1$, hence transformation (4) preserves distribution (1). Therefore, (4) is a *reparameterization* of the distribution. We will refer to the subclass of reparameterizations with $\beta = 0$ as *homogeneous reparameterizations*.

2.2 Mean parameters

The exponential family naturally arises as follows: find a distribution $p(x)$ with maximum entropy and prescribed mean values $\mu \in \mathbb{R}^I$ (a column vector) of the functions ϕ, i.e., $\sum_{x\in X} p(x)\phi(x) = \mu$. Solving this linearly constrained concave maximization task reveals that $p(x)$ must have the form (1), where θ appeared as Lagrange multipliers. Since entropy is strictly concave, μ determines $p(x)$ uniquely. The numbers μ are called the *mean parameters* (or *moments*). Thus, any distribution from the family is uniquely given either by the canonical parameters θ or by the mean parameters μ. Given θ, the corresponding μ can be explicitly computed as $\mu = m(\theta)$ where $m\colon \mathbb{R}^I \to \mathbb{R}^I$ is the map

$$m(\theta) = \sum_{x\in X} p(x\,|\,\theta)\phi(x) = \frac{\sum_{x\in X} \phi(x) \exp\theta\phi(x)}{\sum_{x\in X} \exp\theta\phi(x)} \tag{5}$$

There is no explicit formula for the inverse of the map m. If ϕ is a minimal representation then for any μ there is a single θ satisfying $\mu = m(\theta)$. For an overcomplete representation the ambiguity

[2] We will call θ a *vector*, although strictly formally $\theta \in \mathbb{R}^I$ means that θ is a *mapping* from I to $\mathbb{R}$. For a vector, we should correctly write $\theta \in \mathbb{R}^{|I|}$. This slight inaccuracy will cause no harm.

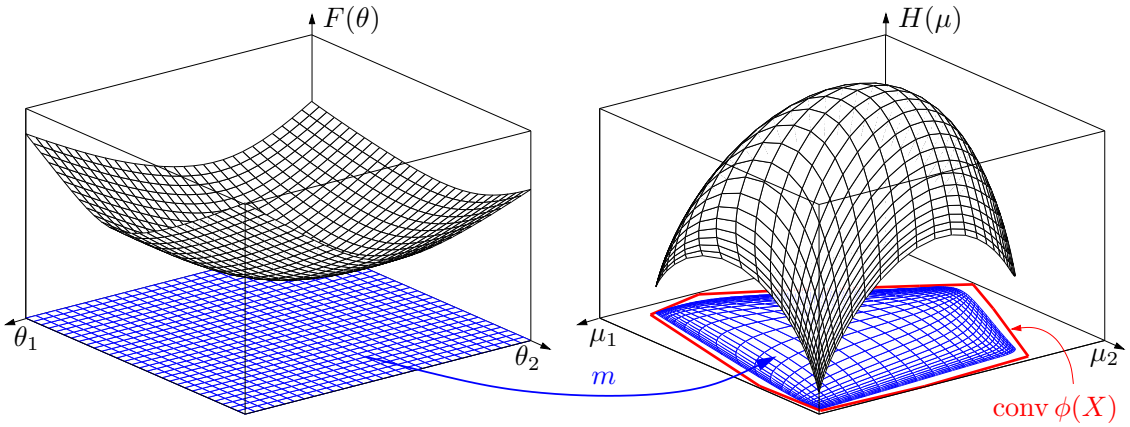


Figure 1: The plot of $F(\theta)$, $m(\theta)$ and $H(\mu)$ for a simple exponential family with $I = \{1, 2\}$ and $|X| = 256$. The the numbers $\phi_i(x)$ were i.i.d. drawn from the normal distribution $\mathcal{N}[0, 1]$.

in solving the equation $\mu = m(\theta)$ is given exactly by reparameterizations, in particular we have $m(\theta + \alpha A + \beta B) = m(\theta)$.

What is the range $m(\mathbb{R}^I)$ of the map m, that is, the set of vectors μ for all possible choices of θ? Let $\phi(X) = \{\, \phi(x) \mid x \in X \,\}$ denote the range of the mapping ϕ, a finite set of vectors from $\mathbb{R}^I$. Obviously, the set of mean value vectors of ϕ realizable by all possible distributions p is the convex hull of $\phi(X)$,

$$\operatorname{conv} \phi(X) = \Big\{ \sum_{x \in X} p(x)\phi(x) \;\Big|\; p\colon X \to \mathbb{R}_+,\ \sum_{x \in X} p(x) = 1 \Big\} \tag{6}$$

Symbol p in (6) denotes all possible distributions over X, not necessarily from the family (1). However, it turns out [12] that almost any element of (6) can be obtained also as the mean of ϕ over a distribution *from* the family (1) – precisely, $m(\mathbb{R}^I)$ is the (relative) interior of $\operatorname{conv} \phi(X)$. The polytope $\operatorname{conv} \phi(X)$ is contained in the affine hull of $\phi(X)$, which, by (3), equals

$$\operatorname{aff} \phi(X) = \{\, \mu \in \mathbb{R}^I \mid A\mu = 0,\ B\mu = 1 \,\} \tag{7}$$

2.3 Entropy

It is easy to obtain by direct calculation that the entropy of distribution (1) as a function of θ is $F(\theta) - \theta m(\theta)$. Let $H(\mu)$ denote the entropy of the distribution from the family as a function of μ. It is defined implicitly: to evaluate $H(\mu)$, we first take θ satisfying $\mu = m(\theta)$ (in case of overcomplete representation, any such θ can be taken) and then let $H(\mu) = F(\theta) - \theta\mu$. The function H is positive and concave and its domain is the relative interior of $\operatorname{conv} \phi(X)$.

Theorem 1. *Any μ from the relative interior of $\operatorname{conv} \phi(X)$ and any θ satisfy*

$$F(\theta) - H(\mu) - \theta\mu \geq 0 \tag{8}$$

where equality holds if and only if $\mu = m(\theta)$, that is, if the distribution defined by θ and the distribution defined by μ are the same.

This theorem can be interpreted in two ways. First, the left-hand side of (8) is the relative entropy (KL-divergence) from a distribution defined by θ to a (generally different) distribution defined by μ. The relative entropy is always non-negative and becomes zero for identical distributions. Second, the functions F and $-H$ are related by convex conjugacy (Legendre-Fenchel transform) [1] and (8) is Fenchel's inequality.

Theorem 1 is consistent with the equality

$$\frac{\mathrm{d}F(\theta)}{\mathrm{d}\theta} = m(\theta) \tag{9}$$

which can be verified also by direct calculation. The minimum of (8) over θ is attained at the stationary point satisfying $\mathrm{d}F(\theta)/\mathrm{d}\theta = \mu$. But at minimum we also have $\mu = m(\theta)$, which yields (9). In case of minimal representation, minimizing (8) over μ yields also the dual equality

$$-\frac{\mathrm{d}H(\mu)}{\mathrm{d}\mu} = m^{-1}(\mu) \tag{10}$$

where m^{-1} denotes the inverse of the map m (which is unique in case of minimal representation).

Figure 1 illustrates the convex conjugacy relation on a simple example.

2.4 Gibbs distribution as an exponential family

Let (V, E) be an undirected graph, where V is a finite set of variables and $E \subseteq \binom{V}{2}$. In the sequel, $N_u = \{\, v \mid \{u, v\} \in E \,\}$ will denote the neighbors and $n_u = |N_u|$ the degree of variable u. Each variable $u \in V$ takes states x_u from a finite domain X_u. Let X be the Cartesian product of the variable domains X_u. Let

$$I = \{\, (u, x_u) \mid u \in V,\ x_u \in X_u \,\} \cup \{\, (uv, x_u x_v) \mid \{u, v\} \in E,\ x_u \in X_u,\ x_v \in X_v \,\}$$

where $(uv, x_u x_v)$ is the same element as $(vu, x_v x_u)$. In accordance with this, the components of vector θ will be denoted $\theta_u(x_u)$ and $\theta_{uv}(x_u, x_v)$. Let $\phi\colon X \to \{0, 1\}^I$ be indicator functions chosen such that

$$\theta\phi(x) = \sum_{u \in V} \theta_u(x_u) + \sum_{\{u,v\} \in E} \theta_{uv}(x_u, x_v) \tag{11}$$

With this choice of (X, I, ϕ), distribution (1) is the *pairwise Gibbs distribution*. The parameters $\mu = m(\theta)$ are the marginals of $p(x|\theta)$ associated with variables V and variable pairs E. The polytope $\operatorname{conv} \phi(X)$ contains all realizable marginal vectors μ and is known as the *marginal polytope*. Moreover, in this case we have $\{0, 1\}^I \cap \operatorname{aff} \phi(X) = \phi(X)$ (which is not true for a general map ϕ).

There are many affine dependencies among functions ϕ. We define matrices A and B indirectly by instantiating expressions (3) and (4). Thus, if we write the homogeneous dependencies $A\mu = 0$ as a set of scalar equalities, we obtain the *marginalization conditions*

$$\sum_{x_v} \mu_{uv}(x_u, x_v) - \mu_u(x_u) = 0 \qquad \forall u \in V,\ v \in N_u,\ x_u \in X_u \tag{12}$$

Similarly, the inhomogeneous dependencies $B\mu = 1$ turn out to be the *normalization conditions*

$$\sum_{x_u} \mu_u(x_u) = 1 \qquad \forall u \in V \tag{13a}$$

$$\sum_{x_u, x_v} \mu_{uv}(x_u, x_v) = 1 \qquad \forall \{u, v\} \in E \tag{13b}$$

where $\mu_u(x_u)$ and $\mu_{uv}(x_u, x_v)$ denote the components of vector μ.

Reparameterization $\theta' = \theta + \alpha A + \beta B$ is in components given by

$$\theta'_u(x_u) \qquad = \theta_u(x_u) \qquad - \sum_{v \in N_u} \alpha_{uv}(x_u) \qquad + \beta_u \tag{14a}$$

$$\theta'_{uv}(x_u, x_v) = \theta_{uv}(x_u, x_v) + \alpha_{uv}(x_u) + \alpha_{vu}(x_v) + \beta_{uv} \tag{14b}$$

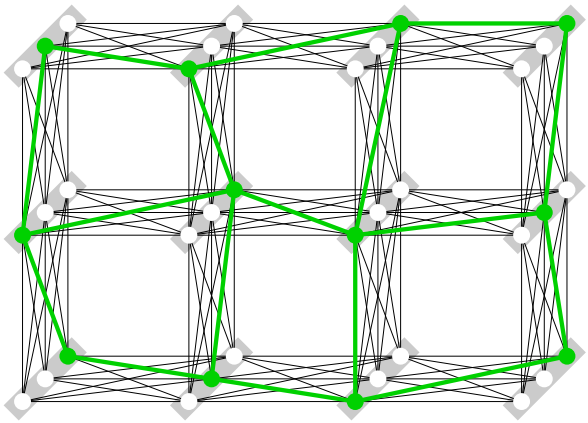


Figure 2: Illustration of a pairwise graphical model with $|X_v| = 3$ states on the 3×4 grid graph (V, E). The grey boxes depict variables and the circles inside them variable states. Set I is formed by all the circles and edges in the figure. Given an assignment $x \in X$, active indicator functions $\phi_i(x)$ are the green nodes and edges and $\theta\phi(x)$ is equal to the sum of numbers θ_i sitting on these green nodes and edges. Set $\phi(X)$ contains all consistent collections of such green nodes and edges.

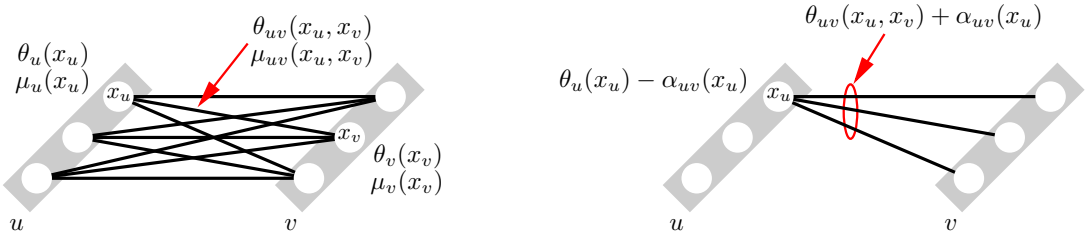


Figure 3: Left: The variables assigned to the nodes end edges of Figure 2. Right: Illustration of homogeneous reparameterization for a single directed edge (u, v) and state x_u.

Let us interpret (14). Homogeneous reparameterization $\theta' = \theta + \alpha A$ can be understood as follows. Suppose we pick a directed edge (u, v) and subtract a unary function $\alpha_{uv}(\cdot)$ from the function $\theta_{uv}(\cdot, \cdot)$ and add the same function to $\theta_u(\cdot)$, that is, we do the transformation

$$\theta_u(x_u) \quad \leftarrow \theta_u(x_u) \quad - \alpha_{uv}(x_u) \tag{15a}$$

$$\theta_{uv}(x_u, x_v) \leftarrow \theta_{uv}(x_u, x_v) + \alpha_{uv}(x_u) \tag{15b}$$

Clearly, this preserves function (11) no matter what $\alpha_{uv}(\cdot)$ is. Composing the elementary reparameterizations (15) for all directed edges explains the terms with α in (14). Pure inhomogeneous reparamerization $\theta' = \theta + \beta B$ means simply adding a constant to each function $\theta_u(\cdot)$ and $\theta_{uv}(\cdot, \cdot)$. In conclusion, function (11) is changed by (14) only by the constant $\beta 1$.

Notice that marginalization+normalization conditions (12)+(13) and reparameterization (14) are 'dual' to each other, via transposing matrices A and B. This was first observed by Schlesinger [9] in LP relaxation of the problem $\max_x \theta\phi(x)$ (see modern revisions [15, 14] of this approach).

The key concepts of this section §2.4 are illustrated in Figures 2 and 3.

2.5 Variational inference

It is of interest in applications to compute the log-partition function $F(\theta)$ and the marginals $m(\theta)$ from given canonical parameters θ. These are examples of *inference* task. These tasks are intractable for Gibbs distribution because the set X is combinatorially large, thus one has to recourse to approximative methods. One of such methods is *variational inference*. It is based on the fact that minimizing (8) over μ allows us to express $F(\theta)$ and $H(\mu)$ of a single distribution in terms of each other,

$$F(\theta) = \max\{\, \theta\mu + H(\mu) \mid \mu > 0,\ \mu \in \operatorname{conv} \phi(X) \,\} \tag{16}$$

This concave maximization problem attains its optimum at a single vector $\mu = m(\theta)$. Thus, rather than calculating $F(\theta)$ and $m(\theta)$ directly, they are evaluated indirectly by solving (16). So far this provides no advantage because both the feasible set $\operatorname{conv}\phi(X)$ and the entropy function H are defined in a combinatorial way. However, it allows us to design methods for approximative inference, by replacing $\operatorname{conv}\phi(X)$ and H with their approximations that have tractable descriptions.

We remark that in statistical mechanics, $-\theta\phi(x)$ resp. $-\theta\mu$ is often refered to as the Gibbs energy and $-\theta\mu - H(\mu)$ as Gibbs free energy [17].

The marginal polytope $\operatorname{conv}\phi(X)$ is typically approximated by the *local polytope* [12]

$$[0,1]^I \cap \operatorname{aff}\phi(X) = \{\, \mu \geq 0 \mid A\mu = 0,\ B\mu = 1 \,\} \tag{17}$$

Clearly, any $\phi\colon X \to [0,1]^I$ satisfies the inclusion $\operatorname{conv}\phi(X) \subseteq [0,1]^I \cap \operatorname{aff}\phi(X)$, hence the local polytope is an outer bound of the marginal polytope. If graph (V,E) is a tree the inclusion becomes equality, for graphs with cycles the inclusion is strict. Therefore, the resulting approximate marginals need not belong to the marginal polytope and they are called *pseudomarginals* or *beliefs*.

Now, instead of (16) we solve the problem

$$\max\{\, \theta\mu + H_{\text{approx}}(\mu) \mid \mu > 0,\ A\mu = 0,\ B\mu = 1 \,\} \tag{18}$$

Its optimal argument and value is an approximation of $m(\theta)$ and $F(\theta)$, respectively.

3 Concave entropy approximation

In this section we revisit variational inference with a concave entropy approximation. Its advantage is that the resulting variational problem (18) is easy to understand using convexity arguments. The reason why we include this section is to contrast it later with the non-concave Bethe entropy approximation.

A number of concave entropy approximations have been proposed [11, 13, 2]. We consider a very simple concave entropy, which on the one hand yields a particularly poor approximation but on the other hand it is easy to deal with and has the same spirit as more complex concave entropies. For this approximation, problem (18) becomes

$$\max\{\, \theta\mu + H_{\text{c}}(\mu) \mid \mu > 0,\ A\mu = 0,\ B\mu = 1 \,\} \tag{19}$$

where

$$H_{\text{c}}(\mu) = -\sum_u \sum_{x_u} \mu_u(x_u) \log \mu_u(x_u) - \sum_{\{u,v\}} \sum_{x_u, x_v} \mu_{uv}(x_u, x_v) \log \mu_{uv}(x_u, x_v) \tag{20}$$

is simply the sum of independent node and edge entropies. The function H_{c} is strictly concave on $\mathbb{R}^I_{++}$ and it is an upper bound on the true entropy H.

3.1 Dual problem

Let us write the Lagrange dual problem to (19). We form the Lagrangian

$$\begin{aligned} L(\mu, \alpha) &= \theta\mu + H_{\text{c}}(\mu) + \alpha A\mu \\ &= (\theta + \alpha A)\mu + H_{\text{c}}(\mu) \end{aligned}$$

The Lagrangian includes the constraint $A\mu = 0$ but not $\mu > 0$ and $B\mu = 1$. Notice that $\theta + \alpha A$ is a homogeneous reparameterization of θ. The dual problem is obtained from the minimax inequality

$$\max_{\mu > 0 \mid B\mu = 1} \min_{\alpha} L(\mu, \alpha) \leq \min_{\alpha} \max_{\mu > 0 \mid B\mu = 1} L(\mu, \alpha) \tag{21}$$

where the left-hand side is equivalent to problem (19) and the right-hand side is the desired dual problem. The dual problem can be written as

$$\min_{\alpha} F_{\text{c}}(\theta + \alpha A) \tag{22}$$

where

$$F_{\mathrm{c}}(\theta) = \max_{\mu>0|B\mu=1} [\theta\mu + H_{\mathrm{c}}(\mu)] \tag{23}$$

Problem (23) can be easily solved, resulting in

$$F_{\mathrm{c}}(\theta) = \sum_u \bigoplus_{x_u} \theta_v(x_v) + \sum_{\{u,v\}} \bigoplus_{x_u,x_v} \theta_{uv}(x_u, x_v) \tag{24}$$

The function F_{c} is obviously convex. It can be shown [16] that it is an upper bound on the true log-partition function F. The dual problem (22) can be interpreted as minimizing the function $F_{\mathrm{c}}(\theta)$ over homogeneous reparameterizations of the original vector θ.

Since the primal problem (19) is concave, (21) holds with equality and we have strong duality.

3.2 Optimality condition

We want to find conditions on which problems (19) and (22) are jointly optimal. Before doing that, we state in Lemma 1 a property of functions F_{c} and H_{c} somewhat analogical to Theorem 1.

We define the map $m_{\mathrm{c}}\colon \mathbb{R}^I \to \mathbb{R}^I$ by $m_{\mathrm{c}}(\theta) = \mu$ where

$$\mu_u(x_u) = \exp\left[\theta_u(x_u) - \bigoplus_{x_u} \theta_u(x_u)\right] \tag{25a}$$

$$\mu_{uv}(x_u, x_v) = \exp\left[\theta_{uv}(x_u, x_v) - \bigoplus_{x_u,x_v} \theta_{uv}(x_u, x_v)\right] \tag{25b}$$

Functions $\mu_u(\cdot)$ and $\mu_{uv}(\cdot,\cdot)$ are positive and normalized, that is, $m_{\mathrm{c}}(\theta) > 0$ and $Bm_{\mathrm{c}}(\theta) = 1$. In fact, they are Gibbs distributions (1) on trivial graphs formed by single node u and single edge $\{u, v\}$, respectively. The numbers μ are often called *pseudomarginals* or *beliefs*. Obviously, $m(\theta + \beta B) = m(\theta)$.

Lemma 1. *For any θ and any $\mu > 0$ satisfying $B\mu = 1$ we have*

$$F_{\mathrm{c}}(\theta) - H_{\mathrm{c}}(\mu) - \theta\mu \geq 0 \tag{26}$$

where the equality holds if and only if $\mu = m_{\mathrm{c}}(\theta)$.

Proof. It is easy to show that the corresponding node and edge terms in (20) and (24) are individually related by convex conjugacy. Then (26) is Fenchel's inequality. ∎

Note that convex conjugacy between F_{c} and $-H_{\mathrm{c}}$ implies that

$$\frac{\mathrm{d}F_{\mathrm{c}}(\theta)}{\mathrm{d}\theta} = m_{\mathrm{c}}(\theta) \tag{27}$$

Theorem 2. *Problems (19) and (22) are jointly optimal if and only if $A\mu = 0$ and $\mu = m_{\mathrm{c}}(\theta+\alpha A)$.*

Proof. Since tasks (19) and (22) are related by strong duality, at their joint optimum their objectives meet. This happens for $\mu > 0$ and θ satisfying $A\mu = 0$, $B\mu = 1$, and $H_{\mathrm{c}}(\mu) + \theta\mu = F_{\mathrm{c}}(\theta + \alpha A)$. Condition $A\mu = 0$ implies that $\theta\mu = (\theta + \alpha A)\mu$, and thus the last equality is the same as

$$F_{\mathrm{c}}(\theta + \alpha A) - H_{\mathrm{c}}(\mu) - (\theta + \alpha A)\mu = 0$$

But this is (26) written for reparameterized θ. The rest follows from Lemma 1. ∎

Eliminating μ from the equations $\mu = m_{\mathrm{c}}(\theta + \alpha A)$ and $A\mu = 0$ yields

$$Am_{\mathrm{c}}(\theta + \alpha A) = 0 \tag{28}$$

Since it no longer involves μ, (28) is the optimality condition for the dual problem (22) only. This result can be obtained in a more direct way, by finding the zero gradient of the dual objective $F_{\mathrm{c}}(\theta + \alpha A)$ with respect to reparameterizations α,

$$\frac{\mathrm{d}F_{\mathrm{c}}(\theta + \alpha A)}{\mathrm{d}\alpha} = A \frac{\mathrm{d}F_{\mathrm{c}}(\theta + \alpha A)}{\mathrm{d}(\theta + \alpha A)} = 0 \tag{29}$$

where the equality in (29) follows from the chain rule. Plugging (27) into (29) yields (28).

Since $\theta + \alpha A$ is a reparameterization of θ, condition (28) has a clear interpretation: *to solve the dual problem (22), we need to reparameterize the original vector θ such that $Am_{\mathrm{c}}(\theta) = 0$.*

3.3 Sum-product diffusion

The obtained optimality condition is the fixed point of an algorithm to solve the dual problem (22). This algorithm [16] is more known in its zero-temperature version as *max-sum diffusion* [6, 15], which solves the LP-relaxation of the problem of finding modes of Gibbs distribution (MAP-MRF) and yields the same bound as several other recent algorithms, most notably TRW-S [5]. Its finite-temperature version is obtained by replacing the operation max with $\oplus$, hence can be called *$\oplus$-sum diffusion*. We can pass to exponential domain, i.e. replace operations $(\oplus, +)$ with $(+, \times)$ and exponentiate all quantities, which yields *sum-product diffusion*. All these 'diffusion' algorithms can be seen as special cases of the marginal consistency algorithm on a commutative semiring [8].

Note that reparameterizations can be done in two ways: either θ is fixed and only α is changed (then the current canonical parameters are $\theta + \alpha A$) or θ itself is changed – we use the latter way.

Definition 1. A vector θ is a fixed point of $\oplus$-sum diffusion if $Am_{\mathrm{c}}(\theta) = 0$.

Let us make the fixed point condition more explicit. Plugging (25) and (12) into $Am_{\mathrm{c}}(\theta) = 0$ yields

$$\theta_u(x_u) - \bigoplus_{x_u} \theta_u(x_u) = \bigoplus_{x_v} \theta_{uv}(x_u, x_v) - \bigoplus_{x_u, x_v} \theta_{uv}(x_u, x_v) \qquad \forall u \in V,\ v \in N_u,\ x_u \in X_u \tag{30}$$

If the graph (V, E) is connected, by a homogeneous reparameterization it can be always achieved that $\bigoplus_{x_u} \theta_u(x_u) = \bigoplus_{x_u, x_v} \theta_{uv}(x_u, x_v)$ holds for every every (u, v), thus these terms cancel out in (30) and we are left with

$$\theta_u(x_u) = \bigoplus_{x_v} \theta_{uv}(x_u, x_v) \qquad \forall u \in V,\ v \in N_u,\ x_u \in X_u \tag{31}$$

This condition says that for every $\{u, v\}$, $\theta_u(x_u)$ and $\theta_v(x_v)$ have to be $\oplus$-marginals of $\theta_{uv}(x_u, x_v)$.

The $\oplus$-sum diffusion update (Algorithm 1) enforces equality (31) on a single directed edge (u, v) by applying homogeneous reparameterization (15) on (u, v). This determines $\alpha_{uv}(\cdot)$ in (15) uniquely. The update is iterated for all directed edges in an arbitrary order. This converges to a state when (31) holds globally. Every update decreases the dual objective $F_{\mathrm{c}}(\theta)$, therefore the algorithm is a block-coordinate descent to minimize $F_{\mathrm{c}}(\theta)$ over homogeneous reparameterizations. Convexity and smoothness of F_{c} guarantees convergence to the global minimum.

Algorithm 1 Update of $\oplus$-sum diffusion on directed edge (u, v).

1: $\forall x_u$: $\alpha_{uv}(x_u) \leftarrow \frac{1}{2}\Big[\theta_u(x_u) - \bigoplus_{x_v} \theta_{uv}(x_u, x_v)\Big]$;
2: $\forall x_u$: $\theta_u(x_u) \leftarrow \theta_u(x_u) - \alpha_{uv}(x_u)$;
3: $\forall x_u, x_v$: $\theta_{uv}(x_u, x_v) \leftarrow \theta_{uv}(x_u, x_v) + \alpha_{uv}(x_u)$;

4 Bethe entropy approximation

In this section we revisit the Bethe variational problem

$$\max\{\,\theta\mu + H_{\mathrm{b}}(\mu) \mid \mu > 0,\ A\mu = 0,\ B\mu = 1\,\} \tag{32}$$

where

$$H_{\mathrm{b}}(\mu) = -\sum_u \sum_{x_u} \mu_u(x_u) \log \mu_u(x_u) - \sum_{\{u,v\}} \sum_{x_u,x_v} \mu_{uv}(x_u,x_v) \log \frac{\mu_{uv}(x_u,x_v)}{\mu_u(x_u)\,\mu_v(x_v)} \tag{33}$$

is the Bethe approximation of the true entropy function H. The negative objective $-\theta\mu - H_{\mathrm{b}}(\mu)$ is known as the Bethe free energy. As pointed out in [12], in the works by Yedidia et al. [18] function H_{b} is used in the alternative form

$$-\sum_u (1-n_u) \sum_{x_u} \mu_u(x_u) \log \mu_u(x_u) - \sum_{\{u,v\}} \sum_{x_u,x_v} \mu_{uv}(x_u,x_v) \log \mu_{uv}(x_u,x_v) \tag{34}$$

Functions (33) and (34) are equal but only on condition $A\mu = 0$. We will further use only the form (33). Unlike convex entropy approximations, H_{b} is neither a lower nor an upper bound on H. In general, function H_{b} is non-concave. If the graph (V,E) is a tree, the Bethe approximation is exact, $H_{\mathrm{b}} = H$, and H_{b} is concave on the set $\{\,\mu > 0 \mid A\mu = 0,\ B\mu = 1\,\}$ but remains non-concave on $\{\,\mu > 0 \mid B\mu = 1\,\}$.

For trees, the local polytope is the marginal polytope and thus (32) exactly equals $F(\theta)$.

4.1 Dual problem

Let us try forming the dual to problem (32). We proceed exactly as in §3.1. The Lagrangian is

$$L(\mu,\alpha) = (\theta + \alpha A)\mu + H_{\mathrm{b}}(\mu)$$

The dual problem is the right-hand side of the minimax inequality

$$\max_{\mu>0|B\mu=1} \min_{\alpha} L(\mu,\alpha) \le \min_{\alpha} \max_{\mu>0|B\mu=1} L(\mu,\alpha) \tag{35}$$

and can be written as

$$\min_{\alpha} \tilde{F}_{\mathrm{b}}(\theta + \alpha A) \tag{36}$$

where

$$\tilde{F}_{\mathrm{b}}(\theta) = \max_{\mu>0|B\mu=1} [\theta\mu + H_{\mathrm{b}}(\mu)] \tag{37}$$

To tackle problem (37) we again use Lagrange multipliers. The solution of (37) must make the derivatives of its Lagrangian $L(\mu,\beta) = (\theta+\beta B)\mu - \beta 1 + H_{\mathrm{b}}(\mu)$ vanish, that is, we solve the system

$$\theta + \beta B + \frac{\mathrm{d}H_{\mathrm{b}}(\mu)}{\mathrm{d}\mu} = 0 \tag{38a}$$

$$B\mu = 1 \tag{38b}$$

for μ and β. The components of the row vector $\mathrm{d}H_{\mathrm{b}}(\mu)/\mathrm{d}\mu$ are

$$\frac{\partial H_{\mathrm{b}}(\mu)}{\partial \mu_u(x_u)} = -\log \mu_u(x_u) - 1 + \frac{1}{\mu_u(x_u)} \sum_{v\in N_u} \sum_{x_v} \mu_{uv}(x_u,x_v) \tag{39a}$$

$$\frac{\partial H_{\mathrm{b}}(\mu)}{\partial \mu_{uv}(x_u,x_v)} = -\log \frac{\mu_{uv}(x_u,x_v)}{\mu_u(x_u)\mu_v(x_v)} - 1 \tag{39b}$$

Plugging this into (38) shows that system (38) has no explicit solution and thus function $\tilde{F}_{\mathrm{b}}$ does not have a closed form. Therefore, we have to give up forming an explicit dual to the Bethe variational problem (32).

Even though the dual does not possess an explicit form, one can ask whether strong duality holds. Of course not, because the function H_{c} is non-concave. Does strong duality hold at least for tree-structured problems? No, because for the minimax inequality (35) to hold with equality, H_{b} would have to be concave on the set $\{\,\mu > 0 \mid B\mu = 1\,\}$ and not only on $\{\,\mu > 0 \mid A\mu = 0,\ B\mu = 1\,\}$.

4.2 Optimality condition

Given the non-existence of an explicit dual to the Bethe variational problem and the absence of strong duality, we will not form the analogy of Theorem 2 for the Bethe entropy approximation. However, a close counterpart to Theorem 2 is the result [18, 17] relating belief propagation fixed points and the Bethe variational problem. We state this result in Theorem 3.

We define the map $m_{\mathrm{b}}\colon \mathbb{R}^I \to \mathbb{R}^I$ by $m_{\mathrm{b}}(\theta) = \mu$ where

$$\mu_u(x_u) = \exp\Big[\theta_u(x_u) - \bigoplus_{x_u} \theta_u(x_u)\Big] \tag{40a}$$

$$\mu_{uv}(x_u, x_v) = \exp\Big[\theta_{uv}(x_u, x_v) + \theta_u(x_u) + \theta_v(x_v) - \bigoplus_{x_u, x_v} [\theta_{uv}(x_u, x_v) + \theta_u(x_u) + \theta_v(x_v)]\Big] \tag{40b}$$

Again, the numbers μ are positive and normalized, $m_{\mathrm{b}}(\theta) > 0$ and $Bm_{\mathrm{b}}(\theta) = 1$. We have also $m(\theta + \beta B) = m(\theta)$.

Theorem 3. *μ is a stationary point of problem (32) if and only if there is α such that $\mu = m_{\mathrm{b}}(\theta + \alpha A)$ and $A\mu = 0$.*

Proof. The Lagrangian of problem (32) reads

$$L(\mu, \alpha, \beta) = \theta\mu + H_{\mathrm{b}}(\mu) + \alpha A\mu + \beta(B\mu - 1)$$

We shall show that (μ, α, β) is a stationary point of $L(\mu, \alpha, \beta)$ for some β if and only if $\mu = m_{\mathrm{b}}(\theta + \alpha A)$ and $A\mu = 0$. We solve the system

$$\frac{\partial L(\mu, \alpha, \beta)}{\partial \mu} = \theta + \alpha A + \beta B + \frac{\mathrm{d}H_{\mathrm{b}}(\mu)}{\mathrm{d}\mu} = 0 \tag{41}$$

Because $A\mu = 0$, in other words $\sum_{x_v} \mu_{uv}(x_u, x_v) = \mu_u(x_u)$, the unpleasant expression (39a) simplifies to

$$\frac{\partial H_{\mathrm{b}}(\mu)}{\partial \mu_u(x_u)} = -\log \mu_u(x_u) - 1 + n_u \tag{42}$$

Now we substitute (42) and (39b) into (41), and then substitute $\mu = m_{\mathrm{b}}(\theta + \alpha A)$ into the result. This leaves us only with terms that do not depend on states. Hence equality (41) can be satisfied by choosing numbers β_u and β_{uv}. ∎

Eliminating μ from $\mu = m_{\mathrm{b}}(\theta + \alpha A)$ and $A\mu = 0$ yields

$$Am_{\mathrm{b}}(\theta + \alpha A) = 0 \tag{43}$$

This shows that to find a stationary point of the Lagrangian of the Bethe variational problem, *we need to reparameterize the original vector θ such that $Am_{\mathrm{b}}(\theta) = 0$.*

Suppose that $Am_{\mathrm{b}}(\theta) = 0$. If our graph is a tree, $m_{\mathrm{b}}(\theta)$ are the exact marginals. If it is a graph with cycles, it trivially follows that $m_{\mathrm{b}}(\theta)$ restricted on any subtree are the exact marginals for this subtree [10].

4.3 Loopy belief propagation

The obtained optimality condition is the fixed point condition of BP. We again assume that reparameterizations are done by changing θ itself rather than α.

Definition 2. A vector θ is a fixed point of belief propagation if $Am_{\mathrm{b}}(\theta) = 0$.

Again, the condition $Am_{\mathrm{b}}(\theta) = 0$ can be made more explicit. Inserting (40) and (12) into it yields

$$\bigoplus_{x_v}[\theta_{uv}(x_u, x_v) + \theta_v(x_v)] = a_{uv} \qquad \forall u \in V,\ v \in N_u,\ x_u \in X_u \tag{44}$$

where $a_{uv} = \bigoplus_{x_u,x_v}[\theta_{uv}(x_u, x_v) + \theta_u(x_u) + \theta_v(x_v)] - \bigoplus_{x_u} \theta_u(x_u)$ are constants independent on x_u. Thus, the condition says that for every directed edge (u, v), the left-hand side of (44) has to be independent on x_u.

There are two versions of belief propagation, with parallel or serial updates – we consider the latter version. Its update, Algorithm 2, enforces equality (44) on a single directed edge (u, v) by applying homogeneous reparameterization (15) on (u, v). This determines $\alpha_{uv}(\cdot)$ in (15) uniquely up to an additive constant. This constant is chosen (on line 2 of the algorithm) so that the values θ stay bounded during the algorithm. The update is iterated for all directed edges in an arbitrary order. Typically this converges to a state when (44) holds globally – however, unlike for $\oplus$-sum diffusion, the convergence is not guaranteed and no quantity is monotonically decreasing or increasing.

Algorithm 2 BP update on directed edge (u, v).

1: $\forall x_u\colon\ \alpha_{uv}(x_u) \leftarrow -\bigoplus_{x_v}[\theta_{uv}(x_u, x_v) + \theta_v(x_v)]$;
2: $\nu \leftarrow \bigoplus_{x_u} \alpha_{uv}(x_u) - \bigoplus_{x_u} 0$; $\quad \forall x_u\colon\ \alpha_{uv}(x_u) \leftarrow \alpha_{uv}(x_u) - \nu$;
3: $\forall x_u\colon\ \theta_u(x_u) \leftarrow \theta_u(x_u) - \alpha_{uv}(x_u)$;
4: $\forall x_u, x_v\colon\ \theta_{uv}(x_u, x_v) \leftarrow \theta_{uv}(x_u, x_v) + \alpha_{uv}(x_u)$;

Let us remark that the zero-temperature version of BP is obtained by replacing $\oplus$ with max in Algorithm 2. Note, then the normalization term $\bigoplus_{x_u} 0 = \log |X_u|$ on line 2 becomes $\max_{x_u} 0 = 0$.

5 New view on belief propagation fixed points

In case of concave entropy approximation, the $\oplus$-sum diffusion algorithm monotonically decreases the smooth convex function $F_{\mathrm{c}}(\theta + \alpha A)$ over reparameterizations α, which guarantees global optimality and convergence. Setting the gradient of this function to zero yields the fixed point conditions of the algorithm, $Am_{\mathrm{c}}(\theta + \alpha A) = 0$.

In contrast, there is no function that would monotonically decrease or increase during the BP algorithm. One could object that this is not true because we have the Bethe variational problem (32), in which we maximize the function $\theta\mu + H_{\mathrm{b}}(\mu)$. But this is of no help because BP does not directly solve problem (32). From the point of view of problem (32), the BP algorithm keeps $\mu = m_{\mathrm{b}}(\theta)$ all the time, which ensures $\mu > 0$ and $B\mu = 1$, and tries to reparameterize θ such that $A\mu = 0$. Therefore, μ is infeasible to (32) until BP has converged and we cannot infer that, for example, the primal objective $\theta\mu + H_{\mathrm{b}}(\mu)$ increases as the BP algorithm proceeds. Of course, the same is true about $\oplus$-sum diffusion.

Seemingly, nothing more can be said about the BP algorithm than it tries to solve the equation system $Am_{\mathrm{b}}(\theta + \alpha A) = 0$ for α. Can we interpret this system as the zero gradient condition of some function of reparameterizations α? In analogy with the equality $F_{\mathrm{c}}(\theta) = \theta m_{\mathrm{c}}(\theta) + H_{\mathrm{c}}[m_{\mathrm{c}}(\theta)]$, which follows from Lemma 1, we can try to obtain such a function as

$$F_{\mathrm{b}}(\theta) = \theta m_{\mathrm{b}}(\theta) + H_{\mathrm{b}}[m_{\mathrm{b}}(\theta)] \tag{45}$$

Plugging (40) and (33) into (45) and simplifying yields

$$F_{\mathrm{b}}(\theta) = \sum_u (1 - n_u) \bigoplus_{x_u} \theta_u(x_u) + \sum_{\{u,v\}} \bigoplus_{x_u,x_v} [\theta_{uv}(x_u, x_v) + \theta_u(x_u) + \theta_v(x_v)] \tag{46}$$

Our main contribution is the observation (in Theorem 4 below) that zero gradients of function F_b with respect to homogeneous reparameterizations are in one-to-one correspondence with BP fixed points. Further in this section we examine the function F_b in detail.

Recalling that there are two forms of the Bethe entropy approximation, (33) and (34), substituting (34) instead of (33) into (45) does not work – zero gradients of the resulting function do not correspond to BP fixed points.

Function F_b is non-convex, due to the negative factors $1 - n_u$. It is no longer convex conjugate with H_b, in particular it is not true that all θ and $\mu > 0$ such that $A\mu = 0$ and $B\mu = 1$ satisfy $F_\mathrm{b}(\theta) - H_\mathrm{b}(\mu) - \theta\mu \geq 0$.

5.1 First-order properties

Let us form the derivative $\mathrm{d}F_\mathrm{b}(\theta)/\mathrm{d}\theta$. It needs some work to obtain its components as

$$\frac{\partial F_\mathrm{b}(\theta)}{\partial \theta_u(x_u)} = \mu_u(x_u) + \sum_{v \in N_u} \gamma_{uv}(x_u) \tag{47a}$$

$$\frac{\partial F_\mathrm{b}(\theta)}{\partial \theta_{uv}(x_u, x_v)} = \mu_{uv}(x_u, x_v) \tag{47b}$$

where $\mu = m_\mathrm{b}(\theta)$ and $\gamma = A\mu$, that is,

$$\gamma_{uv}(x_u) = \sum_{x_v} \mu_{uv}(x_u, x_v) - \mu_u(x_u)$$

Note, comparing this with $\mathrm{d}F_\mathrm{c}(\theta)/\mathrm{d}\theta$ given by (27) we see there is an extra term $\sum_{v \in N_u} \gamma_{uv}(x_u)$.

Next we compute the derivative of $F_\mathrm{b}(\theta + \alpha A)$ with respect to homogeneous reparameterizations α. It can be conveniently computed at point $\alpha = 0$, which is without any loss of generality because the derivative at $\alpha \neq 0$ can be recovered by replacing θ with $\theta + \alpha A$. By the chain rule, we have

$$\left.\frac{\mathrm{d}F_\mathrm{b}(\theta + \alpha A)}{\mathrm{d}\alpha}\right|_{\alpha=0} = A \frac{\mathrm{d}F_\mathrm{b}(\theta)}{\mathrm{d}\theta}$$

which, by (12), means

$$\left.\frac{\partial F_\mathrm{b}(\theta + \alpha A)}{\partial \alpha_{uv}(x_u)}\right|_{\alpha=0} = \sum_{x_v} \frac{\partial F_\mathrm{b}(\theta)}{\partial \theta_{uv}(x_u, x_v)} - \frac{\partial F_\mathrm{b}(\theta)}{\partial \theta_u(x_u)} \tag{48}$$

Substituting (47) into (48) yields

$$\left.\frac{\partial F_\mathrm{b}(\theta + \alpha A)}{\partial \alpha_{uv}(x_u)}\right|_{\alpha=0} = -\sum_{w \in N_u \setminus v} \gamma_{uw}(x_u) \tag{49}$$

Theorem 4. *θ is a BP fixed point if and only if the gradient of $F_\mathrm{b}(\theta)$ with respect to homogeneous reparameterizations vanishes.*

Proof. By Theorem 3, BP fixed point is characterized by system (50a). By (49), zero gradient of $F_\mathrm{b}(\theta + \alpha A)$ is characterized by system (50b).

$$\gamma_{uv}(x_u) = 0 \quad \forall u \in V,\ v \in N_u,\ x_u \in X_u \tag{50a}$$

$$\sum_{w \in N_u \setminus v} \gamma_{uw}(x_u) = 0 \quad \forall u \in V,\ v \in N_u,\ x_u \in X_u \tag{50b}$$

We need to show that these two systems are equivalent. Pick (u, x_u) and write γ_v instead of $\gamma_{uv}(x_u)$ for simplicity. Then we need to show that

$$[\gamma_v = 0 \ \ \forall v \in N_u] \iff \Big[\sum_{w \in N_u \setminus v} \gamma_w = 0 \ \ \forall v \in N_u \Big]$$

One direction is obvious, the other is easily verified. ∎

The current understanding of BP sees it as an iterative algorithm to solve the equation system (50a). Now we have the additional knowledge that the solutions of (50a) are in one-to-one correspondence to zero gradient points of a single function. This zero gradient condition is given by system (50b). In this new view, BP is seen as a block-coordinate search to find a zero gradient point.

Note a subtlety: for solving (50b) we need a slightly different schedule of updates than for solving (50a). To solve system (50a), in each iteration we pick a single directed edge (u, v) and run Algorithm 2 on it. This solves a subset of equations (50a) for the coordinate block $\{\, \alpha_{uv}(x_u) \mid x_u \in X_u \,\}$. However, it does not solve any subset of equations (50b). To solve (50b), in each iteration we need to pick a single node u and run Algorithm 2 on directed edges $\{\, (u, v) \mid v \in N_u \,\}$. This solves a subset of equations (50b) for the coordinate block $\{\, \alpha_{uv}(x_u) \mid v \in N_u,\ x_u \in X_u \,\}$.

5.2 Second-order properties

We form the second derivative (Hessian) of F_b with respect to reparameterizations, i.e., the matrix

$$\frac{\mathrm{d}^2 F_\mathrm{b}(\theta + \alpha A)}{\mathrm{d}\alpha^2}$$

We give the Hessian only on the assumption that $\theta + \alpha A$ is a BP fixed point, where it takes a simpler form. After certain effort (see the appendix), we obtain the elements of the Hessian

$$\frac{\partial^2 F_\mathrm{b}(\theta + \alpha A)}{\partial \alpha_{uv}(x_u)\, \partial \alpha_{u'v'}(x'_{u'})} = \begin{cases} [\mu_u(x_u) - 1]\mu_u(x_u) & \text{if } u' = u,\ v' \neq v,\ x_u = x'_u \\ \mu_u(x_u)\mu_u(x'_u) & \text{if } u' = u,\ v' \neq v,\ x_u \neq x'_u \\ \mu_{uu'}(x_u, x'_{u'}) - \mu_u(x_u)\mu_{u'}(x'_{u'}) & \text{if } \{u, u'\} \in E,\ v' \neq u,\ u' \neq v \\ 0 & \text{otherwise} \end{cases} \tag{51}$$

where $\mu = m_\mathrm{b}(\theta + \alpha A)$.

Theorem 5. *Every zero gradient point of $F_\mathrm{b}(\theta + \alpha A)$ as a function of α is a saddle point.*

Proof. A saddle point means that the Hessian is indefinite at that point. We show that a submatrix of the Hessian is indefinite. Let this submatrix be given by picking u and x_u and setting $u = u'$ and $x_u = x'_u$ in (51). We are left with the matrix with coordinates v and v' with elements

$$\frac{\partial^2 F_\mathrm{b}(\theta + \alpha A)}{\partial \alpha_{uv}(x_u)\, \partial \alpha_{uv'}(x_u)} = \begin{cases} 0 & \text{if } v = v' \\ [\mu_u(x_u) - 1]\mu_u(x_u) & \text{if } v \neq v' \end{cases} \tag{52}$$

Expression (52) takes only two values, depending on whether $v = v'$ or $v \neq v'$. Hence the diagonal elements of the submatrix are zero and all the remaining elements are equal. Such a matrix is inevitably indefinite. ∎

5.3 Invariance properties

Is function F_b invariant to any subclass of reparameterizations? Obviously, $F_\mathrm{b}(\theta + \beta B) = F_\mathrm{b}(\theta) + \beta 1$, hence adding constants β_u and β_{uv} to nodes and edges leaves the function unchanged if $\beta 1 = 0$.

Less obviously, at a BP fixed point the function F_b is invariant to reparameterization on any single directed edge (u, v), as stated by the following theorem.

Theorem 6. *Let θ be a BP fixed point. Let α be such that all its components are zero except the components $\{\, \alpha_{uv}(x_u) \mid x_u \in X_u \,\}$ for a single directed edge (u, v). Then $F_\mathrm{b}(\theta) = F_\mathrm{b}(\theta + \alpha A)$.*

Proof. The described reparameterization affects only node u and edges $\{\, \{u, w\} \mid w \in N_u \setminus v \,\}$. Their contribution of this node and these edges to $F_\mathrm{b}(\theta)$ is

$$(1 - n_u) \bigoplus_{x_u} \theta_u(x_u) + \sum_{w \in N_u \setminus v} \bigoplus_{x_u} \Big[\theta_u(x_u) + \underbrace{\bigoplus_{x_w} [\theta_{uw}(x_u, x_w) + \theta_w(x_w)]}_{a_{uw}} \Big] = \sum_{w \in N_u \setminus v} a_{uw} \tag{53}$$

where the underlined expression equals a constant a_{uw} because θ is a BP fixed point, see (44). The terms $\bigoplus_{x_u} \theta_u(x_u)$ cancel out.

The contribution of the above node and edges to $F_{\mathrm{b}}(\theta + \alpha A)$ is the expression (53) in which $\theta_u(x_u)$ is replaced with $\theta_u(x_u) + \alpha_{uv}(x_u)$. The terms $\bigoplus_{x_u}[\theta_u(x_u) + \alpha_{uv}(x_u)]$ cancel out exactly as in (53) and we are left with the same result. ∎

Thus, at a BP fixed point the function $F_{\mathrm{b}}(\theta + \alpha A)$ is constant along each coordinate block α_{uv}. In fact, it is constant even in several such blocks simultaneously on condition that they do not 'interact'.

6 Discussion

Currently, new insights into BP can be gained either by better understanding BP fixed point equations or by better understanding the Bethe variational problem. We have offered another direction by showing that BP searches for a zero gradient of a single function of reparameterizations, without any constraints.

Our result is elementary – in fact, we merely substituted (40) and (33) into $\theta\mu + H_{\mathrm{b}}(\mu)$. However, this has a clear meaning only via to the link with reparameterizations. To the best of our knowledge, this simple observation was not made before.

On the negative side, all zero gradient points of our function are saddles. One would say that finding a zero gradient point of a single function must be easier than solving a set of equations. But this is not so obvious because finding a saddle point can be much harder than finding a local extreme – especially in our case when we have many variables, in general multiple saddles, and we do not know the shape of the saddles *a priori*.

To be more precise, suppose we have n analytic functions $g_1, \ldots, g_n \colon \mathbb{R}^n \to \mathbb{R}$ and want to find at least one common zero of these functions over $\mathbb{R}^n$. Let us distinguish three cases:

1. The functions g_i are unrelated to each other.
2. The functions g_i are the partial derivatives of some function $f \colon \mathbb{R}^n \to \mathbb{R}$ such that
   (a) all zero gradient points of f are saddle points.
   (b) f is bounded from below and it has at least one local minimum.

Case 2b is clearly easier than case 1 because there are provably convergent algorithms able to find a local minimum. This is indeed the case of concave entropy approximations. We find the following question fundamental: *Is case 2a easier than case 1?* Our preliminary attempts to answer this question are inconclusive.

Though we consider only pairwise Gibbs distributions, we believe the result could be generalized to higher-order versions of loopy BP and cluster variation methods [17]. One could also investigate the meaning of the zero-temperature (max-sum) version of the result, obtained by replacing operation $\oplus$ with max in (44) and (46).

References

[1] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

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

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

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

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

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

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

[8] Tomáš Werner. Marginal consistency: Unifying constraint propagation on commutative semirings. In *Intl. Workshop on Preferences and Soft Constraints (co-located with Conf. on Principles and Practice of Constraint Programming)*, pages 43–57, September 2008.

[9] 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.

[10] 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.

[11] Martin J. Wainwright, Tommi Jaakkola, and Alan S. Willsky. A new class of upper bounds on the log partition function. *IEEE Trans. Information Theory*, 51(7):2313–2335, 2005.

[12] 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.

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

[14] Tomáš Werner. Revisiting the linear programming relaxation approach to Gibbs energy minimization and weighted constraint satisfaction. *IEEE Trans. Pattern Analysis and Machine Intelligence*. To appear in 2010, online preprint available in 2009.

[15] 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.

[16] 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.

[17] 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.

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

A Second derivative of F_{b}

Here we derive the elements of the Hessian (51). We do it by taking the derivative of expression (49) with respect to $\alpha_{u'v'}(x'_{u'})$. In the following expressions we assume that $\mu = m_{\mathrm{b}}(\theta + \alpha A)$ and that all the derivatives are evaluated at $\alpha = 0$.

The derivatives of $\mu_u(x_u)$:

$$\begin{aligned}
\frac{\partial \mu_u(x_u)}{\partial \alpha_{uv}(x_u)} &= -[1-\mu_u(x_u)]\mu_u(x_u) \\
\frac{\partial \mu_u(x_u)}{\partial \alpha_{u'v}(x_{u'})} &= 0 && \text{if } u \neq u' \\
\frac{\partial \mu_u(x_u)}{\partial \alpha_{uv}(x'_u)} &= \mu_u(x'_u)\mu_u(x_u) && \text{if } x_u \neq x'_u
\end{aligned}$$

The derivatives not shown are zero. The derivative of $\mu_{uv}(x_u, x_v)$:

$$\begin{aligned}
\frac{\partial \mu_{uv}(x_u,x_v)}{\partial \alpha_{uv}(x'_u)} &= 0 \\
\frac{\partial \mu_{uv}(x_u,x_v)}{\partial \alpha_{uv'}(x_u)} &= -[1-\mu_{uv}(x_u)]\mu_{uv}(x_u,x_v) && \text{if } v \neq v' \\
\frac{\partial \mu_{uv}(x_u,x_v)}{\partial \alpha_{uv'}(x'_u)} &= \mu_{uv}(x'_u)\mu_{uv}(x_u,x_v) && \text{if } v \neq v',\ x_u \neq x'_u
\end{aligned}$$

where we use the abbreviation $\mu_{uv}(x_u) = \sum_{x_v} \mu_{uv}(x_u, x_v)$. The remaining derivatives are obtained by realizing that $\mu_{uv}(x_u, x_v) = \mu_{vu}(x_v, x_u)$.

The derivative of $\gamma_{uv}(x_u)$:

$$\begin{aligned}
\frac{\partial \gamma_{uv}(x_u)}{\partial \alpha_{uv}(x_u)} &= \sum_{x_v} \frac{\partial \mu_{uv}(x_u,x_v)}{\partial \alpha_{uv}(x_u)} - \frac{\partial \mu_u(x_u)}{\partial \alpha_{uv}(x_u)} = [1-\mu_u(x_u)]\mu_u(x_u) \\
\frac{\partial \gamma_{uv}(x_u)}{\partial \alpha_{uv}(x'_u)} &= \sum_{x_v} \frac{\partial \mu_{uv}(x_u,x_v)}{\partial \alpha_{uv}(x'_u)} - \frac{\partial \mu_u(x_u)}{\partial \alpha_{uv}(x'_u)} = -\mu_u(x_u)\mu_u(x'_u) && \text{if } x_u \neq x'_u \\
\frac{\partial \gamma_{uv}(x_u)}{\partial \alpha_{uv'}(x_u)} &= \sum_{x_v} \frac{\partial \mu_{uv}(x_u,x_v)}{\partial \alpha_{uv'}(x_u)} - \frac{\partial \mu_u(x_u)}{\partial \alpha_{uv'}(x_u)} = [1-\mu_u(x_u)]\mu_u(x_u) - [1-\mu_{uv}(x_u)]\mu_{uv}(x_u) && \text{if } v \neq v' \\
\frac{\partial \gamma_{uv}(x_u)}{\partial \alpha_{uv'}(x'_u)} &= \sum_{x_v} \frac{\partial \mu_{uv}(x_u,x_v)}{\partial \alpha_{uv'}(x'_u)} - \frac{\partial \mu_u(x_u)}{\partial \alpha_{uv'}(x'_u)} = \mu_u(x_u)\mu_u(x'_u) - \mu_{uv}(x_u)\mu_{uv}(x'_u) && \text{if } v \neq v',\ x_u \neq x'_u \\
\frac{\partial \gamma_{uv}(x_u)}{\partial \alpha_{vu}(x_v)} &= \sum_{x_v} \frac{\partial \mu_{uv}(x_u,x_v)}{\partial \alpha_{vu}(x_v)} - \frac{\partial \mu_u(x_u)}{\partial \alpha_{vu}(x_v)} = 0 \\
\frac{\partial \gamma_{uv}(x_u)}{\partial \alpha_{vu'}(x_v)} &= \sum_{x'_v} \frac{\partial \mu_{uv}(x_u,x'_v)}{\partial \alpha_{vu'}(x_v)} - \frac{\partial \mu_u(x_u)}{\partial \alpha_{vu'}(x_v)} = \mu_{uv}(x_u,x_v) - \mu_{uv}(x_u)\mu_{vu}(x_v) && \text{if } u \neq u'
\end{aligned}$$

If θ is a BP fixed point, we have $\mu_{uv}(x_u) = \mu_u(x_u)$ and this simplifies to

$$\begin{aligned}
\frac{\partial \gamma_{uv}(x_u)}{\partial \alpha_{uv}(x_u)} &= [1-\mu_u(x_u)]\mu_u(x_u) \\
\frac{\partial \gamma_{uv}(x_u)}{\partial \alpha_{uv}(x'_u)} &= -\mu_u(x_u)\mu_u(x'_u) && \text{if } x_u \neq x'_u \\
\frac{\partial \gamma_{uv}(x_u)}{\partial \alpha_{uv'}(x_u)} &= 0 && \text{if } v \neq v' \\
\frac{\partial \gamma_{uv}(x_u)}{\partial \alpha_{uv'}(x'_u)} &= 0 && \text{if } v \neq v',\ x_u \neq x'_u \\
\frac{\partial \gamma_{uv}(x_u)}{\partial \alpha_{vu}(x_v)} &= 0 \\
\frac{\partial \gamma_{uv}(x_u)}{\partial \alpha_{vu'}(x_v)} &= \mu_{uv}(x_u,x_v) - \mu_u(x_u)\mu_v(x_v) && \text{if } u \neq u'
\end{aligned}$$

Finally, the derivative of F_{b} with respect to reparameterizations (at a BP fixed point) is given by

$$\frac{\partial^2 F_{\mathrm{b}}(\theta+\alpha A)}{\partial \alpha_{uv}(x_u)^2} = -\sum_{w\in N_u\setminus v} \frac{\partial \gamma_{uw}(x_u)}{\partial \alpha_{uv}(x_u)} = 0$$

$$\frac{\partial^2 F_{\mathrm{b}}(\theta+\alpha A)}{\partial \alpha_{uv}(x_u)\,\partial \alpha_{uv}(x'_u)} = -\sum_{w\in N_u\setminus v} \frac{\partial \gamma_{uw}(x_u)}{\partial \alpha_{uv}(x'_u)} = 0 \qquad \text{if } x_u \neq x'_u$$

$$\frac{\partial^2 F_{\mathrm{b}}(\theta+\alpha A)}{\partial \alpha_{uv}(x_u)\,\partial \alpha_{uv'}(x_u)} = -\sum_{w\in N_u\setminus v} \frac{\partial \gamma_{uw}(x_u)}{\partial \alpha_{uv'}(x_u)} = -[1-\mu_u(x_u)]\mu_u(x_u) \qquad \text{if } v \neq v'$$

$$\frac{\partial^2 F_{\mathrm{b}}(\theta+\alpha A)}{\partial \alpha_{uv}(x_u)\,\partial \alpha_{uv'}(x'_u)} = -\sum_{w\in N_u\setminus v} \frac{\partial \gamma_{uw}(x_u)}{\partial \alpha_{uv'}(x'_u)} = \mu_u(x_u)\mu_u(x'_u) \qquad \text{if } v \neq v',\ x_u \neq x'_u$$

$$\frac{\partial^2 F_{\mathrm{b}}(\theta+\alpha A)}{\partial \alpha_{uv}(x_u)\,\partial \alpha_{vu}(x_v)} = -\sum_{w\in N_u\setminus v} \frac{\partial \gamma_{uw}(x_u)}{\partial \alpha_{vu}(x_v)} = 0$$

$$\frac{\partial^2 F_{\mathrm{b}}(\theta+\alpha A)}{\partial \alpha_{uv}(x_u)\,\partial \alpha_{vu'}(x_v)} = -\sum_{w\in N_u\setminus v} \frac{\partial \gamma_{uw}(x_u)}{\partial \alpha_{vu'}(x_v)} = 0 \qquad \text{if } u \neq u'$$

$$\frac{\partial^2 F_{\mathrm{b}}(\theta+\alpha A)}{\partial \alpha_{uv}(x_u)\,\partial \alpha_{v'u}(x_{v'})} = -\sum_{w\in N_u\setminus v} \frac{\partial \gamma_{uw}(x_u)}{\partial \alpha_{v'u}(x_{v'})} = 0 \qquad \text{if } v' \in N_u$$

$$\frac{\partial^2 F_{\mathrm{b}}(\theta+\alpha A)}{\partial \alpha_{uv}(x_u)\,\partial \alpha_{v'u'}(x_{v'})} = -\sum_{w\in N_u\setminus v} \frac{\partial \gamma_{uw}(x_u)}{\partial \alpha_{v'u'}(x_{v'})} = -\mu_u(x_u)\mu_{v'}(x_{v'}) + \mu_{uv'}(x_u, x_{v'}) \qquad \text{if } v' \in N_u,\ u \neq u'$$

which can be written compactly as (51).

We tested the correctness of expression (51) numerically, by comparing it with the second derivatives obtained by finite differences for small random Gibbs distributions.