BELIEF PROPAGATION ON PARTIALLY ORDERED SETS*

ROBERT J. MCELIECE* AND MUHAMMED YILDIRIM†

Abstract. In this paper, which is based on the important recent work of Yedidia, Freeman, and Weiss, we present a generalized form of belief propagation, viz. *belief propagation on a partially ordered set (PBP)*. PBP is an iterative message-passing algorithm for solving, either exactly or approximately, the *marginalized product density* problem, which is a general computational problem of wide applicability. We will show that PBP can be thought of as an algorithm for minimizing a certain "free energy" function, and by exploiting this interpretation, we will exhibit a one-to-one correspondence between the fixed points of PBP and the stationary points of the free energy.

1. Introduction. This paper, which is based largely on ideas first expounded by Yedidia, Freeman, and Weiss [12–14], introduces a class of iterative message-passing algorithms called *belief propagation on partially ordered sets*, or PBP. PBP includes as special cases ordinary belief propagation [8], probability propagation [9], the generalized distributive law [1, 2], the sum-product algorithm [6], generalized belief propagation [12] (all of these with and without loops), and many other instances whose effectiveness has not yet been investigated in detail. PBP, like all belief-propagation type algorithms, can in principle be used to solve, either exactly or approximately, a wide variety of problems from engineering, physics, and computer science.

Besides giving a careful definition of PBP, we shall argue that even in the presence of loops, PBP does something sensible, namely, it attempts to minimize a certain free energy function. This is a small generalization of the similar theorem stated by YFW in [12]. (When the partially ordered set is loop-free, PBP, as expected, gives exact answers.)

Here is an outline of the paper. In Section 2 we introduce an abstract computational problem, the *marginalized product density* problem, which the PBP algorithm is designed to solve. In Section 3, we give some necessary background on partially ordered sets (posets), and introduce a kind of labelled and numbered poset, a *junction poset*, which is the basic combinatorial structure utilized by PBP. In Section 4, we give the rules for updating and fusing messages on a junction poset that define the PBP algorithm. In Section 5, we review some simple ideas from statistical physics, including the Helmholtz free energy and the variational free energy technique for computing it. In Section 6 we introduce the "Bethe-Kikuchi" techniques which physicists have devised to simplify the variational free energy method for computing the Helmholtz free energy, and we will begin to see the connection between statistical physics and the PBP algorithm.

*This research was supported by NSF grant no. CCR-0118670, and grants from Sony, Qualcomm, and Caltech's Lee Center for Advanced Networking.

†Department of Electrical Engineering, California Institute of Technology, Pasadena, CA 91125, USA.

In Section 7, we will attempt to minimize the Bethe-Kikuchi variational free energy using a set of Lagrange multipliers to enforce an unexpected set of constraints first introduced by Yedidia, Freeman, and Weiss. In Section 8, we prove the main result of the paper, namely, that the PBP fixed points are in one-to-one correspondence with the stationary points of the BK variational free energy. Finally, in Section 9 we summarize our findings and list some important open questions.

(Appendix A contains a theorem which shows that many posets yield BK energy surfaces that are convex; this implies that if the PBP algorithm converges, it will converge to the global minimum of the BK variational free energy. Appendix B contains a proof of the validity of the crucial YFW constraint basis change.)

2. The marginalized product density problem. *In this brief section, after introducing some necessary notation, we define the "marginalized product density" problem, which is the basic computational problem addressed by the poset belief propagation algorithm.*

Let $A = \{0, 1, \dots, q-1\}$ be a finite set with q elements. We represent the elements of A^n as vectors of the form $\boldsymbol{x} = (x_1, x_2, \dots x_n)$, with $x_i \in A$, for $i \in \{1, \dots, n\}$. If $R \subseteq \{1, \dots, n\}$, we denote by A^R the set A^n projected onto the coordinates indexed by R. A typical element of A^R will be denoted by $\boldsymbol{x}_R$. If $p(\boldsymbol{x})$ is a probability density on A^n, $p_R(\boldsymbol{x}_R)$ denotes $p(\boldsymbol{x})$ marginalized onto R, i.e.,

$$p_R(\boldsymbol{x}_R) = \sum_{\boldsymbol{x} \setminus \boldsymbol{x}_R \in A^{R^c}} p(\boldsymbol{x}). \tag{2.1}$$

For example, with $n = 4$ and $R = \{2, 3\}$, we have

$$p_{\{2,3\}}(x_2, x_3) = \sum_{x_1, x_4} p(x_1, x_2, x_3, x_4).$$

If $\mathcal{R} = \{R_1, R_2, \dots, R_M\}$ is a collection of subsets of $\{1, \dots, n\}$, let $\{\boldsymbol{a}_R(\boldsymbol{x}_R)\}_{R \in \mathcal{R}}$ be a family of nonegative "local kernels," i.e., $\boldsymbol{a}_R(\boldsymbol{x}_R)$ is a nonnegative real number for each $\boldsymbol{x}_R \in A^R$. Define the corresponding global probability density function

$$B(\boldsymbol{x}) = \frac{1}{Z} \prod_{R \in \mathcal{R}} \boldsymbol{a}_R(\boldsymbol{x}_R), \tag{2.2}$$

where Z is the appropriate normalization constant, i.e.,

$$Z = \sum_{\boldsymbol{x} \in A^n} \prod_{R \in \mathcal{R}} \boldsymbol{a}_R(\boldsymbol{x}_R). \tag{2.3}$$

Finally, we define the *Helmholtz free energy* as

$$F = -\ln Z. \tag{2.4}$$

(We will have more to say about free energy and the connection between the MPD problem and statistical physics in Sections 5 and 6.)

The corresponding *marginalized product density* (MPD) problem is to compute one or more of the marginal densities

$$B_R(\boldsymbol{x}_R) = \sum_{\boldsymbol{x} \setminus \boldsymbol{x}_R} B(\boldsymbol{x}) = \frac{1}{Z} \sum_{\boldsymbol{x} \setminus \boldsymbol{x}_R} \prod_{S \in \mathcal{R}} \boldsymbol{a}_S(\boldsymbol{x}_S). \tag{2.5}$$

It is the object of this paper to define and study a class of iterative message-passing algorithms for solving, either exactly or approximately, the MPD problem. The data structures on which these messages are passed are called *partially ordered sets.*

3. Posets, labelled posets, and junction posets. *In this section we will give the necessary background on partially ordered sets, including facts about partially ordered sets that are both labelled and numbered, viz. junction posets.*

A finite *partially ordered set*[1] (poset for short) is a finite set P together with a binary relation, denoted $\leq$, which satisfies the following three axioms:

1. For all $\rho \in P$, $\rho \leq \rho$ *(reflexitivity).*
2. If $\rho \leq \sigma$ and $\sigma \leq \rho$, then $\rho = \sigma$ *(antisymmetry).*
3. If $\rho \leq \sigma$ and $\sigma \leq \tau$, then $\rho \leq \tau$ *(transitivity).*

We use the obvious notation $\rho \geq \sigma$ to mean $\sigma \leq \rho$, $\rho < \sigma$ to mean $\rho \leq \sigma$ but $\rho \neq \sigma$, etc. We say two elements ρ and σ are *comparable* if $\rho \leq \sigma$ or $\sigma \leq \rho$; otherwise, ρ and σ are said to be *incomparable.*

If ρ and σ are elements of P such that $\rho > \sigma$, and there is no element τ such that $\rho > \tau > \sigma$, we say that ρ *covers* σ, and write $\rho \triangleright \sigma$ (alternatively, we say that σ is *covered by* ρ and write $\sigma \triangleleft \rho$).

The *Hasse diagram* $H = H(P)$ for P is a graph with vertex set P, with (ρ, σ) being an edge in H if and only if $\rho \triangleright \sigma$. If $e = (\rho, \sigma) \in E$, we call ρ the *initial element* of e and σ the *final element* of E, and write $\operatorname{init} e = \rho$, $\operatorname{fin} e = \sigma$. We represent the ordering of an edge e by placing $\operatorname{init} e$ "above" $\operatorname{fin} e$ in the Hasse diagram. We say that the poset P is *connected* [*treelike*] if H is connected [a tree].

Figure 1 shows the Hasse diagrams for two 9-element posets. In poset (a), for example, $\{\rho : \rho \leq B\} = \{B, E, G, I\}$, whereas in poset (b), $\{\rho : \rho \leq B\} = \{B, G, I\}$.

We assign an *overcounting number* $\phi(\rho)$ to each $\rho \in P$, such that

$$\sum_{\rho : \rho \geq \sigma} \phi(\rho) = 1, \text{ for all } \sigma \in P. \tag{3.1}$$

[1] For more background on posets, see [10, Chapter 3], on which much of this section is based.

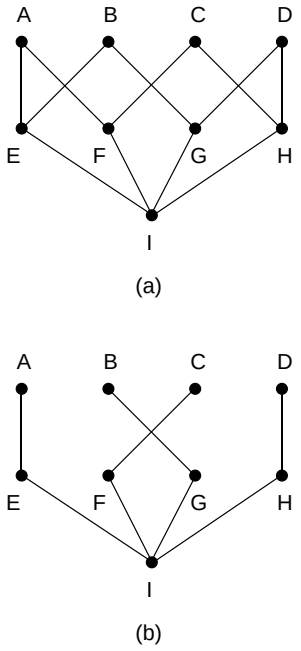


FIG. 1. *The Hasse diagrams for two 9-element posets. Both are connected; (b) is a treelike.*

The numbers $\phi(\rho)$ are integers and are determined uniquely by (3.1). They can be calculated by the following *numbering algorithm*: [2] Number each maximal element $\phi(\rho) = 1$. Then if σ is any unnumbered element of P such that all elements $\rho > \sigma$ are already numbered, define

$$\phi(\sigma) = 1 - \sum_{\rho > \sigma} \phi(\rho).$$

Figure 2 shows the overcounting numbers for the posets in Figure 1.

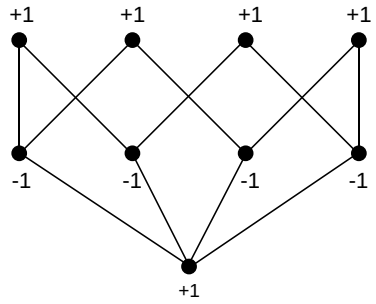
We will also need to label the vertices of the Hasse diagram with subsets of $\{1, \ldots, n\}$, with $L(\rho)$ denoting the label of ρ. We require that such labellings be *order-preserving*, in the sense that

$$L(\sigma) \subseteq L(\rho), \text{ if } \sigma \leq \rho. \tag{3.2}$$

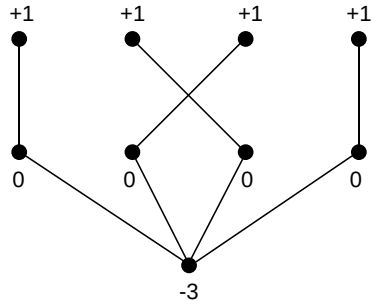
Figure 3 shows an order-preserving labelling of the poset from Figure 1(a).

If P is a labelled poset, and $S \subseteq \{1, \ldots, n\}$, the *$S$-subposet of P* is defined as $P_S = \{\rho : S \subseteq L(\rho)\}$. If, for each $i \in \{1, \ldots, n\}$, the i-subposet

[2] Alternatively, $\phi(\rho) = -\mu(\hat{1}, \rho)$, where $\hat{1}$ is an artificial maximal element added to P, and $\mu(x, y)$ is the Möbius function of the poset $P' = \{\hat{1}, P\}$. It follows that $\phi(\rho) = c_1(\rho) - c_2(\rho) + c_3(\rho) - \cdots$, where $c_i(\rho)$ denotes the number of ascending chains in P of the form $\rho = \rho_1 < \rho_2 < \cdots < p_i$ [10, Chapter 3].



(a)



(b)

FIG. 2. *The Hasse diagrams from Figure 1, with the corresponding overcounting numbers.*

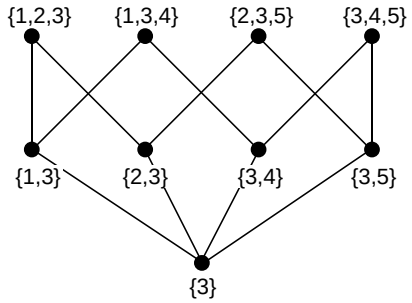


FIG. 3. *The poset from Figure 1(a), with an order-preserving labelling with subsets of* $\{1, 2, 3, 4, 5\}$. *(This is the "natural" CV poset for* $\mathcal{R} = \{\{1, 2, 3\}, \{1, 3, 4\}, \{3, 4, 5\}, \{2, 3, 5\}\}$.*)*

of P is connected, we say that the poset is 1-*connected*. (This is analogous to the junction tree/graph requirements of [1, 2]).[3]

DEFINITION. *A labelled poset which is* 1-*connected is called a* junction poset.

The labelled poset in Figure 3 is a junction poset.

Given a collection $\mathcal{R}$ of subsets of $\{1, \ldots, n\}$ as in Section 2, a junction poset is called a *junction poset for* $\mathcal{R}$ if each element $\rho \in P$ is assigned a subset $\mathcal{R}(\rho) \subseteq \mathcal{R}$, such that:

$$\begin{aligned} \bigcup_{R \in \mathcal{R}(\rho)} R \ &\subseteq L(\rho) \qquad \text{for all } \rho \in P \\ \mathcal{R}(\sigma) \ &\subseteq \mathcal{R}(\rho) \qquad \text{if } \sigma \leq \rho, \end{aligned}$$

and such that the following conditions are satisfied:

$$\sum_{\rho : i \in L(\rho)} \phi(\rho) = 1, \quad \text{for all } i \in \{1, \ldots, n\} \quad (\text{"1-variable balance"}) \tag{3.3}$$

$$\sum_{\rho : R \in \mathcal{R}(\rho)} \phi(\rho) = 1, \quad \text{for all } R \in \mathcal{R} \quad (\text{"conservation of energy"}). \tag{3.4}$$

The idea behind these two important conditions is this. Think of $\phi(\rho)$ as the "multiplicity" of the vertex ρ in the Hasse diagram. Then (3.3) says that the net number of occurrences of each variable x_i, for $i = 1, \ldots, n$ is one; and (3.4) says that the net number of occurrences of each local kernel $\boldsymbol{a}_R(\boldsymbol{x}_R)$ is also one.[4] These conditions will ensure that in the algorithm to be described in Section 4, each piece of "evidence" is counted exactly once.[5]

If we have a junction poset for $\mathcal{R}$, with corresponding local kernels $\{\boldsymbol{a}_R(\boldsymbol{x}_R)\}_{R \in \mathcal{R}}$, the *local kernel at* ρ is defined as follows:

$$\boldsymbol{a}_\rho(\boldsymbol{x}_\rho) = \prod_{R \in \mathcal{R}(\rho)} \boldsymbol{a}_R(\boldsymbol{x}_R),$$

where we have introduced the notation $\boldsymbol{x}_\rho$ as a shorthand for $\boldsymbol{x}_{L(\rho)}$. If $\rho > \sigma$, the *local kernel at* ρ *relative to* σ is similarly defined:

$$\boldsymbol{a}_{\rho \setminus \sigma}(\boldsymbol{x}_\rho) = \prod_{R \in \mathcal{R}(\rho) \setminus \mathcal{R}(\sigma)} \boldsymbol{a}_R(\boldsymbol{x}_R).$$

[3] In general, if for each $S \subseteq \{1, \ldots, n\}$ with $|S| = k$, the S-subposet of P is connected, we say that P is k-connected. Note that if P is treelike, 1-connectivity implies k-connectivity for all k.

[4] The term "conservation of energy" will be justified in Section 5, when we define the local energy functions as $E_R(\boldsymbol{x}_R) = -\ln \boldsymbol{a}_R(\boldsymbol{x}_R)$.

[5] In [7] the authors suggest strengthening the 1-variable balance condition to k-variable balance: $\sum_{\rho : S \in L(\rho)} \phi(\rho) = 1$, for all $S \subseteq \{1, \ldots, n\}$ with $|S| \leq k$ such that P_S is nonempty.

We shall demonstrate in Section 4 that given a junction poset for $\mathcal{R}$, there is a message-passing algorithm on P for solving, exactly or approximately, the corresponding MPD problem. But given $\mathcal{R}$, how can we construct a junction poset for $\mathcal{R}$? We conclude this section by giving three general constructions for junction posets for a given $\mathcal{R}$.

Construction 1. (The junction graph method.) Let $G = (V, E, L)$ be a junction *graph* for $\mathcal{R}$, as described in [2]. Create a junction poset as follows: $P = V \cup E$, with the only relations being of the form $v > e$ if v is one of the vertices of e. The overcounting numbers for this poset are easily seen to be $\phi(v) = 1$ for all $v \in V$, $\phi(e) = -1$ for all $e \in E$. The labels are the same as for the junction graph. One local kernel is assigned to each vertex, i.e., $\mathcal{R}(v_i) = R_i$, for $i = 1, \ldots, M$, but $\mathcal{R}(e) = \emptyset$ for all $e \in E$. This poset clearly satisfies the order-preserving label condition, because the definition of a junction graph requires $L(e) \subseteq L(v_1) \cap L(v_2)$, if $e = (v_1, v_2)$ is an edge. The 1-connectivity and 1-variable balance properties (3.3) hold because of the junction graph condition (the subgraph induced by vertices labelled i form a tree, which necessarily has one more vertex than edge), and the conservation of energy property (3.4) is trivially true, since for each $R \in \mathcal{R}$, there is exactly one vertex v (and no edges) for which $R \in L(v)$. In general, neither k-connectivity nor k-variable balance for $k > 1$ hold. (See Figure 4, which shows a junction graph and the corresponding junction poset. This junction poset is not $\{1, 3\}$-connected and does not satisfy $\{1, 3\}$ variable balance.) ∎

Construction 2. (The cluster variation method [12–14].) Let $\mathcal{S} = \{S_1, \ldots, S_K\}$ be a collection of subsets ("clusters") of $\{1, \ldots, n\}$ such that each R_i is a subset of at least one S_j. Now let P be the poset consisting of all intersections of elements of $\mathcal{S}$, ordered by inclusion. For each $\rho \in P$, we define

$$\begin{aligned} L(\rho) &= \rho \\ R(\rho) &= \{R \in \mathcal{R} : R \subseteq \rho\}. \end{aligned}$$

Here k-connectivity, k-variable balance for all k, and conservation of energy are all automatically satisfied, since for any subset $T \subseteq \{1, \ldots, n\}$, the set of $\rho \in P$ such that $T \subseteq \rho$ is either empty or of the form $\{\rho : \rho \geq \rho(T)\}$, and by (3.1),

$$\sum_{\rho \geq \rho(T)} \phi(\rho) = 1.$$

Applying this for $T = \{i\}$, we get (3.3), and with $T = R \in \mathcal{R}$, we get (3.4). Figure 5 illustrates this construction for $\mathcal{R} = \{\{1, 2\}, \{2, 3\}, \ldots, \{8, 9\}\}$ and

$$\mathcal{S} = \{\{1, 2, 4, 5\}, \{2, 3, 5, 6\}, \{4, 5, 7, 8\}, \{5, 6, 8, 9\}\}.$$

(This example is taken from [14].)

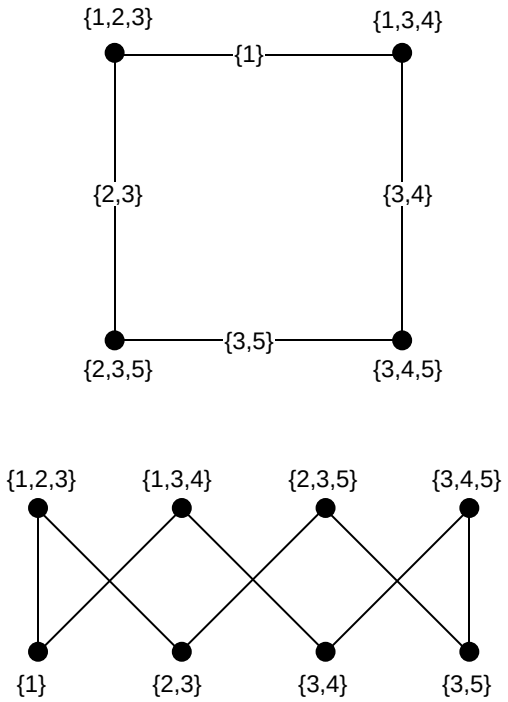


FIG. 4. *A junction graph for* $\mathcal{R} = \{\{1,2,3\},\{1,3,4\},\{2,3,5\},\{3,4,5\}\}$ *(top) and the corresponding junction poset (bottom).*

The special case of this construction when $\mathcal{S} = \mathcal{R} = \{R_1, \ldots, R_M\}$ will be called the "natural" CV method. For example, in Figure 3 we see the "natural" CV junction poset associated with the collection $\mathcal{R} = \{\{1,2,3\},\{1,3,4\},\{3,4,5\},\{2,3,5\}\}$. This should be compared with the junction poset in Figure 4. ∎

Construction 3. (Factor graphs [6], or the Bethe approximation [14].) Given a collection $\mathcal{R} = \{R_1, \ldots, R_M\}$, define a junction poset for $\mathcal{R}$ as follows. For each $j = 1, \ldots, M$, there is a maximal element ρ_j with $L(\rho_j) = R_j$, and $\phi(\rho_j) = 1$. The remaining elements in P are $\{\sigma_1, \ldots, \sigma_n\}$, with $\rho_j > \sigma_i$ iff $i \in R_j$. We define $L(\sigma_i) = \{i\}$, and $\phi(\sigma_i) = -(q_i - 1)$, where q_i is the number of R_j's containing i. Figure 6 illustrates this construction for $\mathcal{R} = \{\{1,2,3\},\{1,3,4\},\{3,4,5\},\{2,3,5\}\}$. In general, junction posets of this type are neither k-connected nor satisfy k-variable balance for $k > 1$. ∎

It is worth noting here that when the "poset-BP" algorithm, which we will describe in the next section, is applied to a junction poset of the type described in Construction 1, the resulting algorithm is equivalent to the "generalized distributive law" algorithm [1]; when it is applied to a poset of the type described in Construction 2, the algorithm is equivalent

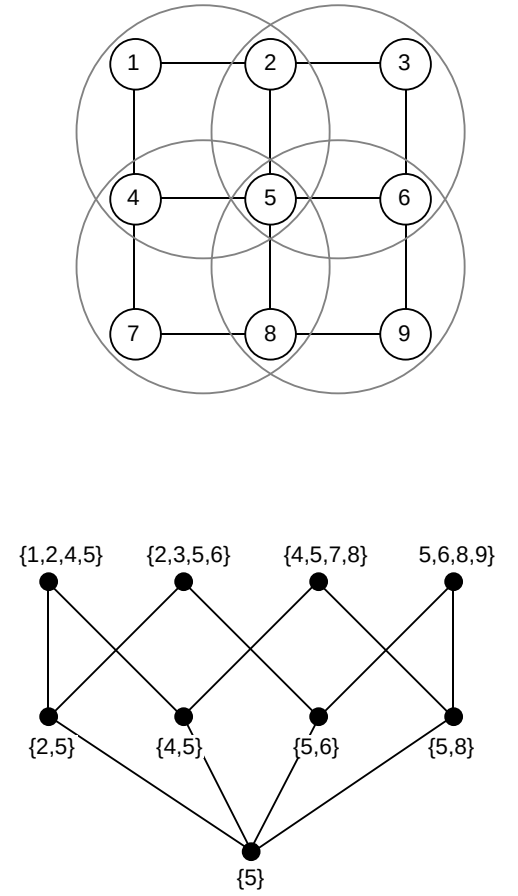


FIG. 5. *Illustrating the cluster variation method. The R_i's are the pairs of variables joined by an edge in the graph at the top. The S_j's (the "clusters") are the sets of variables enclosed by the four circles.*

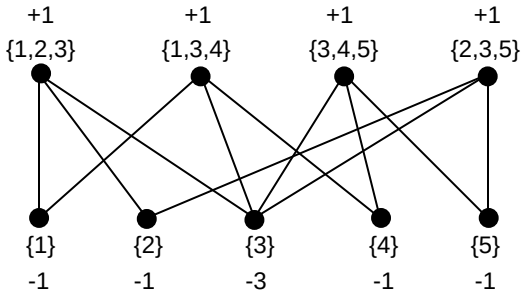


FIG. 6. *Illustrating Construction 3. Here* $\mathcal{R} = \{\{1,2,3\},\{1,3,4\},\{3,4,5\},\{2,3,5\}\}$.

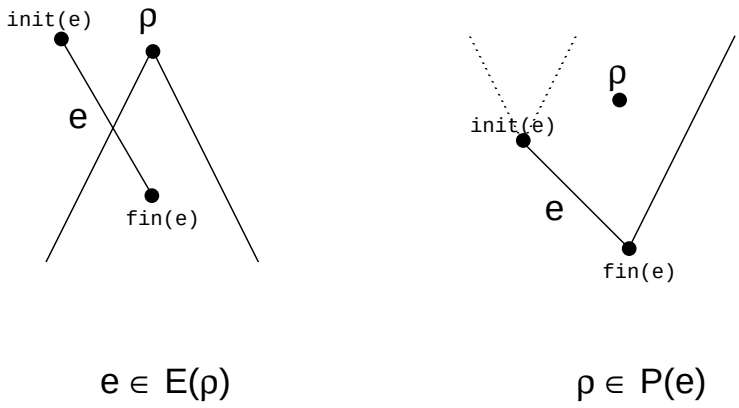


FIG. 7. *Illustrating the definitions of* $E(\rho) = \{e : \text{init}\, e \not\leq \rho, \text{fin}\, e \leq \rho\}$ *and* $P(e) = \{\rho : \rho \not\geq \text{init}\, e,\ \rho \geq \text{fin}\, e\}$.

to "generalized belief propagation" [12]; and for Construction 3, the PBP algorithm is equivalent to the "sum-product" algorithm [6].

4. The poset-BP algorithm. *In this section we will describe an algorithm (the poset-BP algorithm) for solving the MPD problem described in Section 2, assuming we have somehow constructed a junction poset* P *for* $\mathcal{R}$. *As we shall see, the PBP algorithm works by iteratively updating messages on the edges of the Hasse diagram for* P, *and then fusing the messages into beliefs at the vertices.*

In the poset-BP algorithm, associated with each edge $e = (\rho, \sigma) \in E$ is a message m_e. This message is a table of values of a function of the variables $\boldsymbol{x}_\sigma$, and so we write $m_e = \{m_e(\boldsymbol{x}_\sigma)\}$. For example, in the poset shown in Figure 3, the message along the $(\{1,3,4\},\{1,3\})$ edge is a table of the form $f(x_1, x_3)$ for $x_1 \in A$ and $x_3 \in A$.

Each message m_e is sent to one or more elements of P. If $\rho \in P$, the *messages sent to* ρ are $\{m_e : e \in E(\rho)\}$,[6] where

$$E(\rho) = \{e \in E : \text{init}\, e \not\leq \rho,\ \text{fin}\, e \leq \rho\}. \tag{4.1}$$

Conversely, the *destinations for the message* m_e are the elements in the set $P(e)$, where

$$P(e) = \{\rho \in P : \rho \not\geq \text{init}\, e,\ \rho \geq \text{fin}\, e\}. \tag{4.2}$$

See Figure 7 for a pictorial representation of the sets $E(\rho)$ and $P(e)$.

Initially, all entries in the messages are set to 1, and the messages are periodically updated. We will discuss the update rule shortly, but first we

[6] In the terminology of [12], a message m_e with $e \in E(\rho)$ originates "outside" ρ (i.e., $\text{init}\, e \not\leq \rho$) and terminates "inside" ρ (i.e., $\text{fin}\, e \leq \rho$). See Figure 7.

describe the computation of the *beliefs* and the *free energy* associated with a set of messages.

For a given set of messages $\{m_e : e \in E\}$, we define the *belief at* ρ:

$$b_\rho(\boldsymbol{x}_\rho) = \frac{1}{Z_\rho} \boldsymbol{a}_\rho(\boldsymbol{x}_\rho) \prod_{e \in E(\rho)} m_e(\boldsymbol{x}_{\operatorname{fin} e}), \tag{4.3}$$

where the normalization constant Z_ρ is determined so that $\sum_{\boldsymbol{x}_\rho} b_\rho(\boldsymbol{x}_\rho) = 1$, i.e.,

$$Z_\rho = \sum_{\boldsymbol{x}_\rho} \boldsymbol{a}_\rho(\boldsymbol{x}_\rho) \prod_{e \in E(\rho)} m_e(\boldsymbol{x}_{\operatorname{fin} e}). \tag{4.4}$$

The *local free energy* at ρ is defined as

$$F_\rho = -\ln Z_\rho, \tag{4.5}$$

and the *global free energy* is

$$F_P = \sum_{\rho \in P} \phi(\rho) F_\rho. \tag{4.6}$$

How are the messages updated? The update rule is a bit complicated to write down, but easy to motivate. The goal of each message update is to *enforce edge consistency*. By edge consistency we mean the following. If e is an edge with $\operatorname{init} e = \rho$ and $\operatorname{fin} e = \sigma$, we say that e is consistent if the beliefs at ρ and σ, as defined by (4.3), satisfy

$$b_\rho(\boldsymbol{x}_\sigma) = \sum_{\boldsymbol{x}_\rho \setminus \boldsymbol{x}_\sigma} b_\rho(\boldsymbol{x}_\rho) = b_\sigma(\boldsymbol{x}_\sigma) \qquad \text{for all } \boldsymbol{x}_\sigma \in A^{L(\sigma)}. \tag{4.7}$$

In words, (4.7) says that the belief at ρ in $\boldsymbol{x}_\sigma$, obtained by marginalization, agrees with the belief at σ in $\boldsymbol{x}_\sigma$.

The message update rule, then, is this: *Adjust* $m_e(\boldsymbol{x}_\sigma)$ *so that* e *is consistent.* Let us see what this requires. If we substitute the expression on the right side of (4.3) for $b_\rho(\boldsymbol{x}_\rho)$ in (4.7), and substitute the corresponding expression

$$\boldsymbol{a}_\sigma(\boldsymbol{x}_\sigma) \prod_{f \in E(\sigma)} m_f(\boldsymbol{x}_{\operatorname{fin} f})$$

for $b_\sigma(\boldsymbol{x}_\sigma)$ in (4.7), we obtain[7]

$$\sum_{\boldsymbol{x}_\rho \setminus \boldsymbol{x}_\sigma} \left(\boldsymbol{a}_\rho(\boldsymbol{x}_\rho) \prod_{g \in E(\rho)} m_g(\boldsymbol{x}_{\operatorname{fin} g}) \right) = \boldsymbol{a}_\sigma(\boldsymbol{x}_\sigma) \prod_{f \in E(\sigma)} m_f(\boldsymbol{x}_{\operatorname{fin} f}). \tag{4.8}$$

[7] In this calculation, we are free to omit the normalization constants Z_ρ and Z_σ, because multiplying a message by a constant does not affect the local beliefs (b_ρ) or the global free energy F_P. (The local free energies are affected, however.)

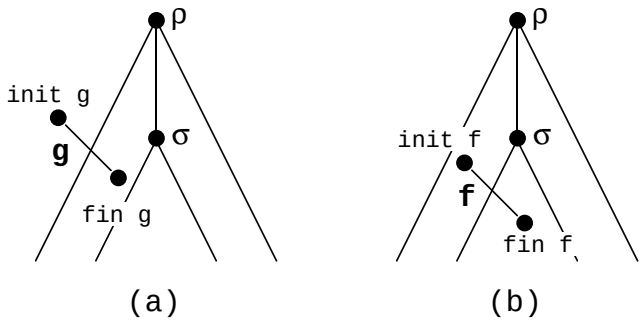


FIG. 8. *Illustrating the sets of edges (a)* $g \in E(\rho)\backslash E(\sigma)$ *and (b)* $f \in E(\sigma)\backslash\{E(\rho)\cup e\}$ *that appear in the numerator resp. denominator of* (4.10).

We can cancel a common factor of

$$\boldsymbol{a}_\sigma(\boldsymbol{x}_\sigma) \prod_{g\in E(\rho)\cap E(\sigma)} m_g(\boldsymbol{x}_{\mathrm{fin}\, g})$$

from both sides of (4.8), thereby obtaining

$$\sum_{\boldsymbol{x}_\rho\backslash\boldsymbol{x}_\sigma} \left(\boldsymbol{a}_{\rho\backslash\sigma}(\boldsymbol{x}_\rho) \prod_{g\in E(\rho)\backslash E(\sigma)} m_g(\boldsymbol{x}_{\mathrm{fin}\, g}) \right) = \prod_{f\in E(\sigma)\backslash E(\rho)} m_f(\boldsymbol{x}_{\mathrm{fin}\, f}). \tag{4.9}$$

Isolating $m_e(\boldsymbol{x}_\sigma)$ from the right side of (4.9) (note that $e \in E(\sigma) \setminus E(\rho)$), we obtain the *message update rule*:

$$m_e(\boldsymbol{x}_\sigma) = \frac{\displaystyle\sum_{\boldsymbol{x}_\rho\backslash\boldsymbol{x}_\sigma} \left(\boldsymbol{a}_{\rho\backslash\sigma}(\boldsymbol{x}_\rho) \prod_{g\in E(\rho)\backslash E(\sigma)} m_g(\boldsymbol{x}_{\mathrm{fin}(g)}) \right)}{\prod_{f\in E(\sigma)\backslash\{E(\rho)\cup e\}} m_f(\boldsymbol{x}_{\mathrm{fin}(f)})}. \tag{4.10}$$

The messages which appear in the numerator and denominator of (4.10) can be understood better if we note that

$$\begin{aligned} E(\rho) \setminus E(\sigma) &= \{g \in E : \mathrm{init}(g) \not\leq \rho, \quad \mathrm{fin}(g) \leq \rho, \not\leq \sigma\} \\ E(\sigma) \setminus E(\rho) &= \{f \in E : \mathrm{init}(f) \leq \rho, \not\leq \sigma, \quad \mathrm{fin}(f) \leq \sigma\} \end{aligned}$$

and refer to Figure 8.

The PBP algorithm proceeds by systematically updating messages according to (4.10).[8] The hope is that eventually, the messages will converge to a fixed point[9] whose associated beliefs are the desired marginals, i.e.,

[8]The scheduling of the message updates is an interesting topic which we will not discuss. However, one reasonable schedule is to update all messages simultaneously, using (4.10), at each iteration of the algorithm.

[9]We say that a set of messages $\{m_e(\boldsymbol{x}_{\mathrm{fin}\, e})\}$ is a *fixed point* of the PBP algorithm if equality holds in (4.10) for all $(e, \boldsymbol{x}_{\mathrm{fin}\, e})$.

$$b_\rho(\boldsymbol{x}_\rho) = B_\rho(\boldsymbol{x}_\rho),$$

(see (2.5)) and whose associated local free energies give the value of the desired Helmholtz free energy, i.e.,

$$F_P = F,$$

where F_P is defined in (4.6). If the poset P is treelike, it can be shown that PBP converges as desired in a finite number of steps. If P is not treelike, i.e., if cycles are present in the Hasse diagram for P, the situation is more complicated, although experiment shows that the beliefs and free energies at attractive fixed points may still give good approximations to the exact values. In any case, however, because of the way we have defined the update rule, we have the following theorem.

THEOREM 1. *If $\{b_\rho(\boldsymbol{x}_\rho)\}_{\rho\in P}$ is the set of beliefs associated with a fixed point of the PBP algorithm, then these beliefs satisfy the edge consistency conditions* (4.7).

It is unfortunate that the update rule (4.10) involves division, in general. We note, however, that some posets forbid configurations of the type shown in Figure 6b, e.g., treelike posets or posets with "depth" ≤ 2. For such posets, the PBP involves only addition and multiplication, which is advantageous in some situations.

In the remainder of the paper we will show that beyond Theorem 1, the fixed points of PBP have a deeper significance. Namely, the PBP fixed points are zero-gradient points on a certain free energy surface, whose global minimum, which is of course among the zero-gradient points, corresponds to the BK approximate solution to the MPD problem. The formal statement of this is Theorem 3 at the end of Section 8.

5. Free energy and the Boltzmann distribution. *In this brief section, we introduce some simple ideas from statistical physics, including the Helmholtz free energy and the variational free energy technique for computing it.*

Imagine a system of n identical particles, each of which can have one of q different "spins" taken from the set $A = \{0, 1, \ldots, q-1\}$. If x_i denotes the spin of the ith particle, we define the state of the system as the vector $\boldsymbol{x} = (x_1, x_2, \ldots x_n)$. In this way, the set A^n can be viewed as a discrete state space S. Now suppose $E(\boldsymbol{x}) = E(x_1, x_2, \ldots, x_n)$ represents the energy of the system (the Hamiltonian) when it is in state $\boldsymbol{x}$. The corresponding *partition function*[10] is defined as

$$Z = \sum_{\boldsymbol{x}\in S} e^{-E(\boldsymbol{x})}, \tag{5.1}$$

[10]In fact, the partition function is also a function of a parameter β, the inverse temperature: $Z = Z(\beta) = \sum_{\boldsymbol{x}\in S} e^{-\beta E(\boldsymbol{x})}$. However, in this paper, we will assume $\beta = 1$, and omit reference to β.

and the *Boltzmann*, or *equilibrium*, density is

$$B(\boldsymbol{x}) = \frac{1}{Z} e^{-E(\boldsymbol{x})} \qquad \text{for } \boldsymbol{x} \in S. \tag{5.2}$$

Finally, the *Helmholtz free energy* of the system is

$$F = -\ln Z.$$

The free energy is of fundamental importance in statistical physics [15, Chapter 2], and physicists have developed a number of ways for calculating it, either exactly or approximately. We will now briefly describe one of these techniques, the *variational free energy* technique.

Suppose $b(\boldsymbol{x})$ represents a "trial" probability of finding the system in state $\boldsymbol{x}$. The corresponding *variational free energy* is defined as

$$\widetilde{F}(b) = U(b) - H(b), \tag{5.3}$$

where $U(b)$ is the average, or internal, energy with respect to the density $b(\boldsymbol{x})$:

$$U(b) = \sum_{\boldsymbol{x} \in S} b(\boldsymbol{x}) E(\boldsymbol{x}), \tag{5.4}$$

and $H(b)$ is the entropy of $b(\boldsymbol{x})$:

$$H(b) = -\sum_{\boldsymbol{x} \in S} b(\boldsymbol{x}) \ln b(\boldsymbol{x}).$$

A routine calculation shows that

$$\widetilde{F}(b) = F + D(b \parallel B),$$

where $D(b \parallel B)$ is the Kullback-Leibler distance between b and B. It then follows from [3, Theorem 2.6.3] that

$$\widetilde{F}(b) \geq F,$$

with equality if and only if $b(\boldsymbol{x}) = B(\boldsymbol{x})$, which is a classical result from statistical physics [11]. In other words,

$$F = \min_{b(\boldsymbol{x})} \widetilde{F}(b) \tag{5.5}$$

$$B(\boldsymbol{x}) = \operatorname*{argmin}_{b(\boldsymbol{x})} \widetilde{F}(b). \tag{5.6}$$

6. The Bethe-Kikuchi approximation to the variational free energy. *In this section we will introduce the "Bethe-Kikuchi" techniques which physicists have devised to simplify the variational free energy method for computing the Helmholtz free energy. We will begin to see the connection between statistical physics and the PBP algorithm.*

According to (5.5), one method for computing the Helmholtz free energy F is to use calculus to minimize the variational free energy $\widetilde{F}(b)$ over all densities $b(\boldsymbol{x})$. However, this involves minimizing a function of the q^n variables $\{b(\boldsymbol{x}) : \boldsymbol{x} \in A^n\}$, which is not an attractive prospect, unless q^n is quite small.

One way around this problem is to approximate the variational free energy $\widetilde{F}(b)$ with a simpler function of b, say $\widetilde{F}_{\text{simple}}(b)$, and then to minimize $\widetilde{F}_{\text{simple}}(b)$, thereby obtaining an approximation to F of the form

$$F \approx \min_{b(\boldsymbol{x})} \widetilde{F}_{\text{simple}}(b).$$

The *Bethe-Kikuchi cluster variation method*, which we will now describe, is a method of this type.

For the BK method to work, it is necessary that the energy function can be decomposed in the form

$$E(\boldsymbol{x}) = \sum_{R \in \mathcal{R}} E_R(\boldsymbol{x}_R), \tag{6.1}$$

where $\mathcal{R}$ is a collection of subsets of $\{1, \ldots, n\}$, as in Sections 2–4. If this is the case, and if P is a junction poset for $\mathcal{R}$, the idea of the BK method is to distribute the energy function $E(\boldsymbol{x})$ around P by defining "local Hamiltonians" $E_\rho(\boldsymbol{x}_\rho)$ as

$$E_\rho(\boldsymbol{x}_\rho) = \sum_{R \in \mathcal{R}(\rho)} E_R(\boldsymbol{x}_R),$$

and to make an approximation to the variational free energy $\widetilde{F}(b)$ of the form $\widetilde{F}(b) \approx \widetilde{F}_P(b)$, where

$$\widetilde{F}_P(b) = \sum_{\rho \in P} \phi(\rho) \widetilde{F}_\rho(b_\rho),$$

where $\widetilde{F}_\rho(b_\rho)$, the *variational free energy at* ρ, is defined as the difference between the *internal energy at* ρ and the *entropy at* ρ.[11] Thus (cf. (5.3))

$$\widetilde{F}_\rho(b_\rho) = U_\rho(b_\rho) - H_\rho(b_\rho),$$

[11] Here $b_\rho(\boldsymbol{x}_\rho)$, is the ρ-marginal belief, defined as $b_\rho(\boldsymbol{x}_\rho) = \sum_{\boldsymbol{x} \setminus \boldsymbol{x}_\rho} b(\boldsymbol{x})$.

where

$$U_\rho(b_\rho) = \sum_{x_\rho} b_\rho(\boldsymbol{x}_\rho) E_\rho(\boldsymbol{x}_\rho),$$

and

$$H(b_\rho) = -\sum_{x_\rho} b_\rho(\boldsymbol{x}_\rho) \ln b_\rho(\boldsymbol{x}_\rho).$$

Because of the "conservation of energy" property (3.4), we find, after a short calculation,

$$U(b) = \sum_{\rho \in P} \phi(\rho) U_\rho(b_\rho).$$

This might lead us to hope that the local entropies might behave in the same way, i.e.,

$$H(b) \stackrel{?}{=} \sum_{\rho \in P} \phi(\rho) H(b_\rho), \tag{6.2}$$

in which case $\widetilde{F}_P(b)$ would be exactly equal to $\widetilde{F}(b)$. However, this hope is in vain. For example, using the junction poset of Figure 3, the hope (6.2) becomes

$$\begin{aligned} H(X_1, X_2, X_3, X_4, X_5) \stackrel{?}{=}\ & H(X_1, X_2, X_3) + H(X_1, X_3, X_4) \\ & + H(X_2, X_3, X_5) + H(X_3, X_4, X_5) \\ & - H(X_1, X_3) - H(X_2, X_3) \\ & - H(X_3, X_4) - H(X_3, X_5) + H(X_3), \end{aligned} \tag{6.3}$$

which is false in general. However, the junction poset "variable balance" property (3.3) guarantees that each X_i is counted exactly once on the right side of (6.2), so it is plausible that the "$\stackrel{?}{=}$" in (6.2) can be replaced with "$\approx$."

In any case, *BK free energy* with respect to the junction poset P is defined as follows:

$$F_P = \operatorname*{cmin}_{\{b_\rho\}} \widetilde{F}_P(\{b_\rho\}), \tag{6.4}$$

where in (6.4) the "cmin" means that the minimum is constrained by the the following conditions on the "marginal" variables $b_\rho(\boldsymbol{x}_\rho)$, for $\rho \in P$ and $\boldsymbol{x}_\rho \in A^{L(\rho)}$:

$$\sum_{\boldsymbol{x}_\rho} b_\rho(\boldsymbol{x}_\rho) = 1, \qquad \text{for all } \rho \in P. \tag{6.5}$$

$$\sum_{\boldsymbol{x}_\rho \setminus \boldsymbol{x}_\sigma} b_\rho(\boldsymbol{x}_\rho) = b_\sigma(\boldsymbol{x}_\sigma), \quad \text{for all } (\rho, \sigma) \in E \text{ and all } \boldsymbol{x}_\sigma \in A^{L(\sigma)}. \tag{6.6}$$

The BK "approximate Boltzmann" beliefs are then the optimizing marginals:

$$\{B_\rho^P\} = \operatorname*{argcmin}_{\{b_\rho\}} \widetilde{F}_P(\{b_\rho\}),$$

The hope, of course, is that F_P will be a good approximation to the Helmholtz free energy F and that the $B_\rho^P(\boldsymbol{x}_\rho)$ will be a good approximation to the Boltzmann marginals $B_\rho(\boldsymbol{x}_\rho)$.

7. Minimizing $\widetilde{F}_P$. *In this section, we will attempt to minimize $\widetilde{F}_P(b)$, using standard techniques from calculus. We begin by using a set of Lagrange multipliers to enforce the edge consistency constraints, but then unexpectedly switch to a different set of constraints, an idea first introduced by Yedidia, Freeman, and Weiss.*

To minimize $\widetilde{F}_P(\{b_\rho(\boldsymbol{x}_\rho)\})$ (see (6.4)), subject to the constraints (6.5) and (6.6), it is natural to set up the following Lagrangian:

$$\begin{aligned}\mathcal{L}_0 \;&= \widetilde{F}_P(\{b_\rho(\boldsymbol{x}_\rho)\}) + \sum_{e\in E}\sum_{\boldsymbol{x}_{\mathrm{fin}\,e}} \lambda(e,\boldsymbol{x}_{\mathrm{fin}\,e})g(e,\boldsymbol{x}_{\mathrm{fin}\,e}) \\ &\quad + \sum_{\rho\in P}\theta_\rho\Bigl(\sum_{\boldsymbol{x}_\rho} b_\rho(\boldsymbol{x}_\rho)\Bigr),\end{aligned} \tag{7.1}$$

where the θ_ρ's are Lagrange multipliers which enforce the constraints (6.5), and the $\{\lambda(e,\boldsymbol{x}_{\mathrm{fin}\,e})\}$'s are Lagrange multipliers which enforce the constraints (6.6), i.e.,

$$g(e,\boldsymbol{x}_{\mathrm{fin}\,e}) = b_{\mathrm{init}\,e}(\boldsymbol{x}_{\mathrm{fin}\,e}) - b_{\mathrm{fin}\,e}(\boldsymbol{x}_{\mathrm{fin}\,e}) = 0. \tag{7.2}$$

We shall call the constraints (7.2) *edge consistency* (cf. (4.7)). This approach is not unreasonable,[12] but a subtler approach, due to Yedidia, Freeman, and Weiss [12–14], yields neater results.

The YFW trick is to minimize $\widetilde{F}_P(\{b_\rho(\boldsymbol{x}_\rho)\})$ not with respect to the edge constraints $g(e,\boldsymbol{x}_{\mathrm{fin}\,e}) = 0$ in (7.2), but rather with respect to a different, but as we shall see ultimately equivalent, set of "weak" edge constraints, viz., $f(e,\boldsymbol{x}_{\mathrm{fin}\,e}) = 0$, where

$$f(e,\boldsymbol{x}_{\mathrm{fin}\,e}) = \sum_{\rho\in P(e)} \phi(\rho) b_\rho(\boldsymbol{x}_{\mathrm{fin}\,e}), \tag{7.3}$$

with the set $P(e)$ as defined in (4.2).

Note that each belief that appears in (7.3) is a belief about $\boldsymbol{x}_{\mathrm{fin}\,e}$, and edge consistency requires all such terms to be equal. But for any edge e,

$$\begin{aligned}\sum_{\rho\in P(e)} \phi(\rho) \;&= \sum_{\rho\geq \mathrm{fin}\,e} \phi(\rho) - \sum_{\rho\geq \mathrm{init}\,e} \phi(\rho) \\ &= 1 - 1 = 0,\end{aligned} \tag{7.4}$$

[12] Indeed, for the restricted class of Eulerian posets [10, Section 3.8], this straightforward approach leads to a PBP algorithm which does not involve division.

so that edge consistency implies $f(e, \boldsymbol{x}_{\text{fin}\, e}) = 0$, i.e., weak edge consistency. The following theorem shows that provided P is connected and $\phi(\rho)$ is never 0, the converse is true.

THEOREM 2. *If P is connected and $\phi(\rho) \neq 0$ for all $\rho \in P$, then weak edge consistency implies edge consistency.*

Proof. The proof of Theorem 2 is given in Appendix B. ∎

With respect to the new edge constraints (the f's), the appropriate Lagrangian for minimizing $\widetilde{F}_P(b)$ is

$$\begin{aligned} \mathcal{L}_1 \;=\; & \widetilde{F}_P(\{b_\rho\}) \\ & + \sum_{e \in E} \sum_{\boldsymbol{x}_{\text{fin}\, e}} \mu(e, \boldsymbol{x}_{\text{fin}\, e}) \sum_{\rho \in P(e)} \phi(\rho) \Bigg(\sum_{\boldsymbol{x}_\rho \setminus \boldsymbol{x}_{\text{fin}\, e}} b_\rho(\boldsymbol{x}_\rho) \Bigg) \\ & + \sum_{\rho \in P} \theta_\rho \Bigg(\sum_{\boldsymbol{x}_\rho} b_\rho(\boldsymbol{x}_\rho) \Bigg), \end{aligned} \tag{7.5}$$

where $\{\mu(e, \boldsymbol{x}_{\text{fin}\, e})\}$ are Lagrange multipliers which enforce the weak edge constraints (7.3). The stationary points of $\mathcal{L}_1$ are the points for which all the partial derivatives vanish. But for $\rho \in P$ and $\boldsymbol{x}_\rho \in A^{L(\rho)}$, we have

$$\frac{\partial \mathcal{L}_1}{\partial b_\rho(\boldsymbol{x}_\rho)} = \phi(\rho) \left(E_\rho(\boldsymbol{x}_\rho) + 1 + \ln b_\rho(\boldsymbol{x}_\rho) \right) + \phi(\rho) \sum_{e \in E(\rho)} \mu(e, \boldsymbol{x}_{\text{fin}\, e}) + \theta_\rho,$$

which equals zero (since $\phi(\rho) \neq 0$) iff

$$\ln b_\rho(\boldsymbol{x}_\rho) = -\Bigg(E_\rho(\boldsymbol{x}_\rho) - F_\rho + \sum_{e \in E(\rho)} \mu(e, \boldsymbol{x}_{\text{fin}\, e}) \Bigg), \tag{7.6}$$

where $F_\rho = -(1 + \theta_\rho / \phi(\rho))$. In short, $\{b_\rho(\boldsymbol{x}_\rho)\}$ is a stationary point for $\widetilde{F}_P(b)$ if and only if there exist constants F_ρ and $\mu(e, \boldsymbol{x}_{\text{fin}\, e})$ such that (7.6) holds.

8. Proof of equivalence of PBP fixed points and BK stationary points. *In this section we prove the main result of the paper, namely, that the PBP fixed points are in one-to-one correspondence with the stationary points of $\widetilde{F}_P(b)$.*

Thus let $\{m_e(\boldsymbol{x}_{\text{fin}\, e})\}$ be a PBP fixed point. Let $\{b_\rho(\boldsymbol{x}_\rho)\}$ be the corresponding set of beliefs, defined by (4.3). These beliefs must satisfy edge consistency, by Theorem 1. Now define a set of Lagrange multipliers by

$$\begin{aligned} \mu(e, \boldsymbol{x}_{\text{fin}\, e}) &= -\ln m_e(\boldsymbol{x}_{\text{fin}\, e}) \\ F_\rho &= -\ln Z_\rho, \end{aligned}$$

and set of local Hamiltonians by

$$E_R(\boldsymbol{x}_R) = -\ln \boldsymbol{a}_R(\boldsymbol{x}_R).$$

If we take the logarithm of both sides of (4.3), we obtain an equation identical to (7.6). Thus *given any PBP fixed point we can produce a unique* $\widetilde{F}_P(b)$ *stationary point.*

Conversely, suppose $\{b_\rho(\boldsymbol{x}_\rho)\}$, $\{\mu(e, \boldsymbol{x}_{\text{fin}\, e})\}$ is a stationary point of $\widetilde{F}_P(b)$. Define a set of messages by

$$m_e(\boldsymbol{x}_{\text{fin}\, e}) = e^{-\mu(e, \boldsymbol{x}_{\text{fin}\, e})},$$

and a set of local kernels by

$$\boldsymbol{a}_R(\boldsymbol{x}_R) = e^{-E_R(\boldsymbol{x}_R)}.$$

Then if we exponentiate (7.6), we reproduce (4.3). Since the beliefs satisfy the edge consistency conditions, by following the derivation that led from (4.3) to (4.9) and (4.10), we conclude that $\{m_e(\boldsymbol{x}_{\text{fin}\, e})\}$ is a PBP fixed point. Thus *given any* $\widetilde{F}_P(b)$ *stationary point we can produce a unique PBP fixed point.*

We conclude by summarizing our findings in a theorem.

THEOREM 3. *The beliefs* $\{b_\rho(\boldsymbol{x}_\rho)\}_{\rho \in P}$ *associated with the fixed points of the PBP algorithm, with respect to the local kernels* $\{\boldsymbol{a}_R(\boldsymbol{x}_R)\}$, *are in one-to-one correspondence with the stationary points of the BK variational free energy* $\widetilde{F}_P(\{b_\rho\})$, *with respect to the energy functions defined by*

$$E_R(\boldsymbol{x}_R) = -\ln \boldsymbol{a}_R(\boldsymbol{x}_R).$$

Furthermore, the value of $\widetilde{F}_P(\{b_\rho\})$ *at one of these fixed points is given by*

$$\widetilde{F}_P = \sum_{\rho \in P} \phi(\rho) F_\rho,$$

where the local free energy F_ρ *is that associated with the fixed point (see* (4.5)*).*

9. Summary and conclusions. In this paper, which is based on the recent work of Yedidia, Freeman, and Weiss, we have presented a generalized form of belief propagation, viz. *belief propagation on a partially ordered set (PBP).* PBP is an iterative message-passing algorithm for solving, either exactly or approximately, the *marginalized product density* problem, which is a general computational problem of wide applicability. PBP is exact if the Hasse diagram for the underlying poset has no cycles, but when the Hasse diagram has cycles, its performance, while often quite good, is difficult to predict. However, again using ideas of YFW, we have shown that the PBP algorithm can also be thought of as an algorithm for minimizing a certain "Bethe-Kikuchi free energy" function. By exploiting this interpretation of PBP, we were able to exhibit a one-to-one relationship between the fixed points of PBP and the stationary points of the BK variational free energy. While this result leaves much unexplained, it does

at least show that PBP is doing something sensible, even in the presence of loops.

We conclude with a list of important open questions.

• For a given MPD problem, or rather for a given collection $\mathcal{R}$ of subsets of $\{1, 2, \ldots, n\}$, how can one find all corresponding junction posets? More ambitiously, how can one find the "best" junction poset for a given problem?

• Can Theorem 3 be strengthened to show that the *stable* fixed points of PBP correspond to the *local minima* of the BK variational free energy?

• Under what circumstances does the PBP algorithm converge?

• Under what circumstances can we guarantee that each step in the PBP leads to a decrease in the BK variational free energy? (This would guarantee that PBP would converge to a local minimum of $\widetilde{F}_P(b)$.)

• What is the relationship between the BK approximate free energy and the exact, or Helmholtz, free energy?

• Can other combinatorial optimization methods, e.g. simulated annealing, be used to minimize the BK variational free energy, thereby leading to alternative "BP" algorithms?

APPENDIX

A. On the convexity of $\widetilde{F}_P(b)$. *In this section we will show that for some posets P, the BK variational free energy $\widetilde{F}_P(b)$ is convex $\cup$, and therefore has a unique stationary point, which is a global minimum. In this favorable situation, Theorem 3 guarantees that if the PBP algorithm converges, it will converge to a set of beliefs corresponding to the Bethe-Kikuchi free energy F_P.*

Let $f(\rho)$ be a real-valued function defined on a poset P. We say that $f(\rho)$ is *monotone* if $\rho \leq \sigma$ implies $f(\rho) \leq f(\sigma)$. If the poset P has the property that

$$\sum_{\rho \in P} \phi(\rho) f(\rho) \geq 0$$

for all nonnegative monotone functions $f(\rho)$, we say that P *has property C*.

For example, any treelike poset (e.g. Figure 1(b)), or any poset whose Hasse diagram has exactly one loop (e.g. Figure 4(b)), has property C. Many other posets, e.g. the one depicted in Figure 1(a), and those in Figure 9, also enjoy property C. (The poset in Figure 6, however, does not have Property C.)

THEOREM 4. *If the poset P has property C, then the variational free energy $\widetilde{F}_P(b)$ is convex $\cup$.*

The proof of Theorem 4 depends on the following facts from calculus.

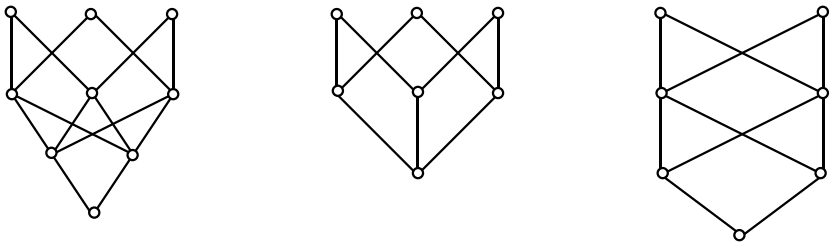


FIG. 9. *Three posets that have Property C. (These were obtained by "beheading" three of the "Eulerian" posets in Fig. 3-30 in [10].)*

Let K be a convex set in R^n. If $\boldsymbol{x} \in K$, a *K-legal direction* from $\boldsymbol{x}$ is a vector $\boldsymbol{y}$ such that $\boldsymbol{x} + \epsilon \boldsymbol{y} \in K$ for all sufficiently small $\epsilon > 0$. We denote the set of K-legal directions from $\boldsymbol{x}$ by $L_K(\boldsymbol{x})$.

Suppose $f(\boldsymbol{x})$ is a real-valued function defined on K. We say that $f(\boldsymbol{x})$ is *convex* $\cup$ on K if, for any $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ in K,

$$f(\epsilon \boldsymbol{x}_1 + (1-\epsilon)\boldsymbol{x}_2) \leq \epsilon f(\boldsymbol{x}_1) + (1-\epsilon) f(\boldsymbol{x}_2),$$

for all $0 < \epsilon < 1$. If $f(\boldsymbol{x})$ is twice differentiable, we can test for convexity with calculus. To this end, define the *Hessian* of $f(\boldsymbol{x})$ for $\boldsymbol{x} \in K$ as the following $n \times n$ matrix:

$$[H(\boldsymbol{x})]_{i,j} = \frac{\partial^2 f(\boldsymbol{x})}{\partial x_i \partial x_j},$$

and let $Q(\boldsymbol{x}, \boldsymbol{y})$ be the corresponding quadratic form:

$$Q(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{y} H(\boldsymbol{x}) \boldsymbol{y}^T = \sum_{i,j} [H(\boldsymbol{x})]_{i,j} y_i y_j.$$

THEOREM 5. *$f(\boldsymbol{x})$ is convex $\cup$ on K if and only if $Q(\boldsymbol{x}, \boldsymbol{y}) \geq 0$ for all $\boldsymbol{x} \in K$ and all $\boldsymbol{y} \in L_K(\boldsymbol{x})$.*

Proof. See Fleming [4]. ∎

In our application of Theorem 5, we wish to test the convexity of the variational free energy with respect to a poset P. This free energy, denoted by $\widetilde{F}_P$, is a function of the variables $\{b_\rho(\boldsymbol{x}_\rho)\}$, and is defined as

$$\text{(A.1)} \qquad \widetilde{F}_P(\{b_\rho(\boldsymbol{x}_\rho)\}) = \sum_{\rho \in P} \phi(\rho) \Bigg(\sum_{\boldsymbol{x}_\rho} b_\rho(\boldsymbol{x}_\rho) E_\rho(\boldsymbol{x}_\rho) + \sum_{\boldsymbol{x}_\rho} b_\rho(\boldsymbol{x}_\rho) \ln b_\rho(\boldsymbol{x}_\rho) \Bigg).$$

The constraints on the variables, which define the convex set K, are

$$\text{(A.2)} \qquad \sum_{\boldsymbol{x}_\rho \backslash \boldsymbol{x}_\sigma} b_\rho(\boldsymbol{x}_\rho) = b_\sigma(\boldsymbol{x}_\sigma) \qquad \text{for all edges } (\rho, \sigma) \in E,$$

and

$$\sum_{\boldsymbol{x}_\rho} b_\rho(\boldsymbol{x}_\rho) = 1 \qquad \text{for all } \rho \in P. \tag{A.3}$$

The legal directions from a point $\{b_\rho(\boldsymbol{x}_\rho)\}$ are all $\{y_\rho(\boldsymbol{x}_\rho)\}$ satisfying

$$\sum_{\boldsymbol{x}_\rho \setminus \boldsymbol{x}_\sigma} y_\rho(\boldsymbol{x}_\rho) = y_\sigma(\boldsymbol{x}_\sigma) \qquad \text{for all edges } (\rho, \sigma) \in E, \tag{A.4}$$

and

$$\sum_{\boldsymbol{x}_\rho} y_\rho(\boldsymbol{x}_\rho) = 0 \qquad \text{for all } p \in P. \tag{A.5}$$

The Hessian of $\widetilde{F}_P$ is easily seen to be a diagonal matrix with $b_\rho(\boldsymbol{x}_\rho)$ diagonal entry $\phi(\rho)/b_\rho(\boldsymbol{x}_\rho)$, so that the corresponding quadratic form is

$$Q(b, y) = \sum_{\rho \in P} \phi(\rho) \sum_{\boldsymbol{x}_\rho} \frac{y_\rho(\boldsymbol{x}_\rho)^2}{b_\rho(\boldsymbol{x}_\rho)}. \tag{A.6}$$

The question now is this: Is $Q(b, y)$, as defined by (A.6), nonnegative for all b's satisfying (A.2) and (A.3), and all y's satisfying (A.4) and (A.5)? The following Lemma will help us answer this question.

LEMMA 1. *The function*

$$f(\rho) = \sum_{\boldsymbol{x}_\rho} \frac{y_\rho(\boldsymbol{x}_\rho)^2}{b_\rho(\boldsymbol{x}_\rho)}$$

is nonnegative monotone on P, for any legal direction $\{y_\rho(\boldsymbol{x}_\rho)\}$ from the belief set $\{b_\rho(\boldsymbol{x}_\rho)\}$.

Proof. If e is an edge of P with $\operatorname{init} e = \rho$, $\operatorname{fin} e = \sigma$, Schwarz's inequality applied to (A.2) and (A.4) implies

$$\sum_{\boldsymbol{x}_\rho \setminus \boldsymbol{x}_\sigma} \frac{y_\rho(\boldsymbol{x}_\rho)^2}{b_\rho(\boldsymbol{x}_\rho)} \;\geq\; \frac{y_\sigma(\boldsymbol{x}_\sigma)^2}{b_\sigma(\boldsymbol{x}_\sigma)}. \tag{A.7}$$

Summing both sides of (A.7) over all $\boldsymbol{x}_\sigma \in A^{L(\sigma)}$, we obtain

$$\sum_{\boldsymbol{x}_\rho} \frac{y_\rho(\boldsymbol{x}_\rho)^2}{b_\rho(\boldsymbol{x}_\rho)} \;\geq\; \sum_{\boldsymbol{x}_\sigma} \frac{y_\sigma(\boldsymbol{x}_\sigma)^2}{b_\sigma(\boldsymbol{x}_\sigma)} \tag{A.8}$$

which says that the (nonnegative) function $f(\rho)$ is monotone on P. ∎

The proof of Theorem 4 is now immediate, since if P has property C, it follows from Lemma 1 that

$$\sum_{\rho \in P} \phi(\rho) \sum_{\boldsymbol{x}_\rho} \frac{y_\rho(\boldsymbol{x}_\rho)^2}{b_\rho(\boldsymbol{x}_\rho)} \;\geq\; 0,$$

i.e., $Q(b, y) \geq 0$, which proves $\widetilde{F}_P$ is convex. ∎

B. Proof of Theorem 2. *In this appendix, we give our proof of Theorem 2. For convenience, we begin by repeating the statement of the theorem.*

THEOREM 2. *If P is connected and $\phi(\rho) \neq 0$ for all $\rho \in P$, then weak edge consistency implies edge consistency.*

Proof. It is easy to show that if P is connected, edge consistency, i.e., $g(e, \boldsymbol{x}_{\text{fin}\, e}) = 0$ for all $(e, \boldsymbol{x}_{\text{fin}\, e})$ holds if and only if

$$b_\rho(\boldsymbol{x}_\sigma) = b_\sigma(\boldsymbol{x}_\sigma) \qquad \text{for all} \quad \rho \geq \sigma. \tag{B.1}$$

Therefore it is enough to show that if $f(e, \boldsymbol{x}_{\text{fin}\, e}) = 0$ for all $(e, \boldsymbol{x}_{\text{fin}\, e})$, then the beliefs likewise satisfy (B.1). We will do this by induction on the *level* of σ, where the level of an element $\rho \in P$ is defined to be the maximal length of a maximal chain from ρ to a maximal element. Thus the maximal elements have level zero, and inductively,

$$\text{level}(\sigma) = 1 + \max_{\rho : \rho > \sigma} \text{level}(\rho).$$

If $\text{level}(\sigma) = 0$, there is nothing to prove, so we begin with $\text{level}(\sigma) = 1$. If $\text{level}(\sigma) = 1$, let $\rho_1, \ldots, \rho_K$ be the elements at level 0 such that $\rho_i > \sigma$. Then $\phi(\sigma) = -(K-1)$, and with $e = (\rho_i, \sigma)$, the condition $f(e, \boldsymbol{x}_{\text{fin}\, e}) = 0$ becomes

$$-(K-1) b_\sigma(\boldsymbol{x}_\sigma) + \sum_{\substack{j=1 \\ j \neq i}}^{K} b_{\rho_j}(\boldsymbol{x}_\sigma) = 0, \qquad \text{for} \quad i = 1, \ldots, K. \tag{B.2}$$

Eq. (B.2) is a set of K linear equations in the K unknowns b_σ, b_{ρ_j} $(j \neq i)$, whose unique solution is $b_{\rho_j}(\boldsymbol{x}_\sigma) = b_\sigma(\boldsymbol{x}_\sigma)$ for $j = 1, \ldots, K$, which completes our proof of (B.1) for $\text{level}(\sigma) = 1$.

If $\text{level}(\sigma) = L \geq 2$, define

$$Q = Q(\sigma) = \{\rho : \rho > \sigma\},$$

and let $Q_1, \ldots, Q_K$ be the connected components of Q, so that we have

$$Q = \biguplus_{i=1}^{K} Q_i \quad \text{(disjoint union)}.$$

Then by the induction hypothesis, for each $i = 1, \ldots, K$, there exists a function $y_i(\boldsymbol{x}_\sigma)$ such that

$$b_\rho(\boldsymbol{x}_\sigma) = y_i(\boldsymbol{x}_\sigma) \quad \text{for all} \quad \rho \in Q_i.$$

Now define

$$G_i = \sum_{\rho \in Q_i} \phi(\rho),$$

so that

$$\phi(\rho) = 1 - \sum_{i=1}^{K} G_i.$$

Since we are assuming $\phi(\rho) \neq 0$, it follows that

$$\sum_{i=1}^{K} G_i \neq 1.$$

Now let us choose a minimal element $\rho' \in Q_i$, and apply the condition $f(e, \boldsymbol{x}_{\text{fin}\, e}) = 0$ to the edge $e = (\rho', \sigma)$. We have (defining $y_j = y_j(\boldsymbol{x}_\sigma)$ and $b = b_\sigma(\boldsymbol{x}_\sigma)$):

$$\begin{aligned} f(e, \boldsymbol{x}_\sigma) &= \sum_{\rho \geq \sigma} \phi(\rho) b_\rho(\boldsymbol{x}_\sigma) - \sum_{\rho \geq \rho'} \phi(\rho) b_\rho(\boldsymbol{x}_\sigma) \\ &= \phi(\sigma) b_\sigma(\boldsymbol{x}_\sigma) + \sum_{\rho > \sigma} \phi(\rho) b_\rho(\boldsymbol{x}_\sigma) - y_i \sum_{\rho \geq \rho'} \phi(\rho) \\ &= \phi(\sigma) b_\sigma(\boldsymbol{x}_\sigma) + \sum_{j=1}^{K} \sum_{\rho \in Q_j} \phi(\rho) b_\rho(\boldsymbol{x}_\sigma) - y_i \\ &= \phi(\sigma) b + \sum_{j=1}^{K} y_j \sum_{\rho \in Q_j} \phi(\rho) - y_i \\ &= \phi(\sigma) b + \sum_{j=1}^{K} G_j y_j - y_i \end{aligned}$$

Thus we have the following K simultaneous linear equations in the unknowns $y_1, \ldots, y_K$:

$$\sum_{j=1}^{K} G_j y_j - y_i = -\phi(\sigma) b, \quad \text{for } i = 1, \ldots, K. \tag{B.3}$$

In matrix-vector notation, this is

$$(\boldsymbol{g}^T \boldsymbol{u} - I_K)\boldsymbol{y} = -\phi(\sigma) b \boldsymbol{u},$$

where $\boldsymbol{g} = (G_1, G_2, \ldots, G_K)$, $\boldsymbol{u} = (1, 1, \ldots, 1)$, and $\boldsymbol{y} = (y_1, \ldots, y_K)$. But the matrix $(\boldsymbol{g}^T \boldsymbol{u} - I_K)$ is nonsingular (since the only nonzero eigenvalue of $\boldsymbol{g}^T \boldsymbol{u}$ us $\boldsymbol{g}\boldsymbol{u}^T = \sum_j G_j \neq 1$),[13] and so the solution to (B.3) is unique. But with $y_j = b$ for $j = 1, \ldots, K$, each of the K equations in (B.3) becomes

$$\left(\sum_{j=1}^{K} G_j - 1 \right) b = -\phi(\sigma) b,$$

[13] See e.g. [5, Prob. 1.4.1], where it is shown that the only nonzero eigenvalue of a rank one matrix $\boldsymbol{a}^T \boldsymbol{b}$ is $\boldsymbol{a}\boldsymbol{b}^T$.

which is correct, since $\phi(\sigma) = 1 - \sum_{j=1}^{K} G_j$. Thus we have proved that $y_j = b$ for all $j = 1, \ldots, K$, i.e.,

$$b_\rho(\boldsymbol{x}_\sigma) = b_\sigma(\boldsymbol{x}_\sigma) \quad \text{for all } \rho > \sigma,$$

which completes the inductive proof of Theorem 2. ∎

REFERENCES

[1] S.M. Aji and R.J. McEliece, "The generalized distributive law," *IEEE Trans. Inform. Theory*, **46**(2): 325–343 (March 2000).

[2] S.M. Aji and R.J. McEliece, "The generalized distributive law and free energy minimization," Proc. 2001 Allerton Conf. Comm. Control and Computing (Oct. 2001).

[3] T.M. Cover and J.A. Thomas, *Elements of Information Theory*. New York: John Wiley and Sons, 1991.

[4] W.H. Fleming, *Functions of Several Variables*. Reading, Mass.: Addison-Wesley, 1965.

[5] R.A. Horn and C.R. Johnson, *Matrix Analysis*. Cambridge: Cambridge University Press, 1985.

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

[7] P. Pakzad and V. Anantharam, "Belief propagation and statistical physics," *Proc. 2002 Conf. Inform. Sciences and Systems*, Princeton U., March 2002.

[8] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*. San Francisco: Morgan Kaufmann, 1988.

[9] G.R. Shafer and P.P. Shenoy, "Probability propagation," *Ann. Mat. Art. Intell.*, **2**: 327–352 (1990).

[10] R.P. Stanley, *Enumerative Combinatorics, Vol. I.* (Cambridge Studies in Advanced Mathematics 49) Cambridge: Cambridge University Press, 1997.

[11] J.S. Yedidia, "An idiosyncratic journey beyond mean field theory," pp. 21–35 in *Advanced Mean Field Methods, Theory and Practice*, eds. Manfred Opper and David Saad, MIT Press, 2001.

[12] J.S. Yedidia, W.T. Freeman, and Y. Weiss, "Generalized belief propagation," pp. 689–695 in *Advances in Neural Information Processing Systems 13* (2000) eds. Todd K. Leen, Thomas G. Dietterich, and Volker Tresp.

[13] J.S. Yedidia, W.T. Freeman, and Y. Weiss, "Bethe free energy, Kikuchi approximations, and belief propagation algorithms," available at `www.merl.com/papers/TR2001-16/`

[14] J.S. Yedidia, W.T. Freeman, and Y. Weiss, "Constructing free energy approximations and generalized belief propagation algorithms," available at `www.merl.com/papers/TR2002-35/`

[15] J.M. Yeomans, *Statistical Mechanics of Phase Transitions*. Oxford: Oxford University Press, 1992.