The Generalized Distributive Law and Free Energy Minimization*

Srinivas M. Aji
Rainfinity, Inc.
87 N. Raymond Ave. Suite 200
Pasadena, CA 91103
saji@rainfinity.com

Robert J. McEliece
Department of Electrical Engineering
California Institute of Technology
Pasadena, CA 91125
rjm@systems.caltech.edu

Abstract.

In an important recent paper, Yedidia, Freeman, and Weiss [7] showed that there is a close connection between the belief propagation algorithm for probabilistic inference and the Bethe-Kikuchi approximation to the variational free energy in statistical physics. In this paper, we will recast the YFW results in the context of the "generalized distributive law" [1] formulation of belief propagation. Our main result is that if the GDL is applied to junction graph, the fixed points of the algorithm are in one-to-one correspondence with the stationary points of a certain Bethe-Kikuchi free energy. If the junction graph has no cycles, the BK free energy is convex and has a unique stationary point, which is a global minimum. On the other hand, if the junction graph has cycles, the main result at least shows that the GDL is trying to do something sensible.

1. Introduction.

The goals of this paper are twofold: first, to obtain a better understanding of iterative, belief propagation (BP)-like solutions to the general probabilistic inference (PI) problem when cycles are present in the underlying graph $\mathcal{G}$; and second, to design improved iterative solutions to PI problems. The tools we use are also twofold: the "generalized distributive law" formulation of BP [1], and the recent results of Yedidia, Freeman, and Weiss [6,7], linking the behavior of BP algorithms to certain ideas from statistical physics.

If $\mathcal{G}$ hs no cycles, it is well known that BP converges to an exact solution to the inference problem in a finite number of steps [1, 3]. But what if $\mathcal{G}$ has cycles? Experimetally, BP is often seen to work well in this situation, but there is little theoretical understanding of why this should be so. In this paper, by recasting the YFW results, we shall see (Theorem 2, Section 5) that if the GDL converges to a given set of "beliefs," these beliefs correspond to a zero gradient point (conjecturally a local minimum) of the "Bethe-Kikuchi free energy," which is a function whose global minimum is an approximation to the solution to the inference problem. In short, even $\mathcal{G}$ has cycles, the GDL still does something sensible. (If $\mathcal{G}$ has no cycles, the BK free energy is convex and the only stationary point is a global minimum,, which is an exact solution to the inference problem.) Since we also show that a given PI problem can typically be represented by many different junction graphs (Theorem 1 and its Corollary, Section 2), this suggests that there can be many good iterative solutions to the problem. We plan to explore these possibilities in a follow-up paper.

* This work was supported by NSF grant no. CCR-9804793, and grants from the Sony Corp., Qualcomm, and Caltech's Lee Center for Advanced Networking.

Here is an outline of the paper. In Section 2, we introduce the notion of a junction graph, which is essential for the entire paper. In Section 3, we describe a generic PI problem, and the "generalized distributive law," which is an iterative, BP-like algorithm for solving the PI problem by passing messages on a junction graph. In Section 4 we state and prove a fundamental theorem from statistical physics, viz., that the variational free energy is always greater than or equal to the free energy, with equality if and only if the system is in Boltzmann equilibrium. In Section 5, we show that the PI problem introduced in Section 3 can be recast as a free energy computation. We then show how this free energy computation can be simplified by using an approximation, the Bethe-Kikuchi approximation (which is also based on a junction graph), to the variational free energy. Finally, we prove our main result, viz., the fixed points of the GDL are in one-to-one correspondence with the stationary points of the Bethe-Kikuchi free energy (Theorem 2, Section 5).

2. Junction Graphs.

Following Stanley [4], we let $[n] = \{1, 2, \ldots, n\}$. An $[n]$-*junction graph* is a labelled, undirected graph $\mathcal{G} = (V, E, L)$, in which each vertex $v \in V$ and each edge $e \in E$ is labelled with a subset of $[n]$, denoted by $L(v)$, and $L(e)$, respectively. If $e = \{v_1, v_2\}$ is an edge joining the vertices v_1 and v_2, we require that

$$L(e) \subseteq L(v_1) \cap L(v_2).$$

Furthermore, we require that for each $k \in [n]$, the subgraph of $\mathcal{G}$ consisting only of the vertices and edges which contain k in their labels, is a tree. Figures 1 through 4 give examples of junction graphs.

If $\mathcal{R} = \{R_1, \ldots, R_M\}$ is a collection of subsets of $[n]$, we say that an $[n]$-junction graph $\mathcal{G} = (V, E, L)$ is a *junction graph for* $\mathcal{R}$ if $\{L(v_1), \ldots, L(v_M)\} = \mathcal{R}$. In [1] it was shown that there is not always a junction *tree* for an arbitrary $\mathcal{R}$. The situation for junction graphs is more favorable, as the following theorem shows.

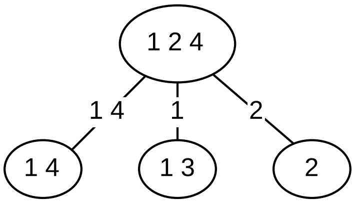

Figure 1. A junction graph for $\mathcal{R} = \{\{1, 2, 4\}, \{1, 3\}, \{1, 4\}, \{2\}\}$. (This is a junction tree.)

Theorem 1. *For any collection $\mathcal{R} = \{R_1, \ldots, R_M\}$ of subsets of $[n]$, there is a junction graph for $\mathcal{R}$.*

Proof: Begin with a complete graph $\mathcal{G}$ with vertex set $V = \{v_1, \ldots, v_M\}$, vertex labels $L(v_i) = R_i$, and edge labels $L(v_i, v_j) = R_i \cap R_j$. For each $k \in [n]$, let $\mathcal{G}_k = (V_k, E_k)$ be the subgraph of $\mathcal{G}$ consisting of those vertices and edges whose label contains k.

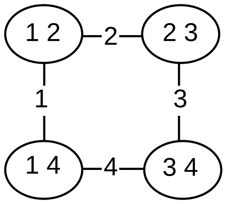

Figure 2. A junction graph for
$\mathcal{R} = \{\{1, 2, \}, \{2, 3\}, \{3, 4\}, \{1, 4\}\}$.

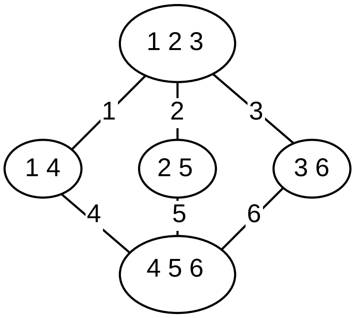

Figure 3. A junction graph for
$\mathcal{R} = \{\{1, 2, 3\}, \{1, 4\}, \{2, 5\}, \{3, 6\}, \{4, 5, 6\}\}$.

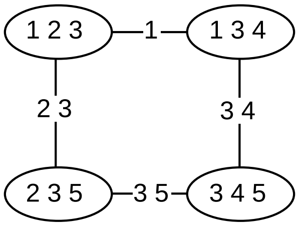

Figure 4. A junction graph for
$\mathcal{R} = \{\{1, 2, 3\}, \{1, 3, 4\}, \{2, 3, 5\}, \{3, 4, 5\}\}$.)

Clearly $\mathcal{G}_k$ is a complete graph, since if $k \in L(v_i) = R_i$ and $k \in L(v_j) = R_j$, then $k \in L(v_i) \cap L(v_j) = R_i \cap R_j$. Now let T_k be any spanning tree of $\mathcal{G}_k$, and delete k from the labels of all edges in E_k except those in T_k. The resulting labelled graph is a junction graph for $\mathcal{R}$. ▪

Using the fact that a complete graph on m vertices has exactly m^{m-2} spanning trees [4, Prop. 5.3.2], we have the following corollary.

Corollary. *If m_i denotes the number of sets R_j such that $i \in R_j$, then the number of junction graphs for $\mathcal{R}$ is $\prod_{i=1}^{n} m_i^{m_i - 2}$.*

3. Probabilistic Inference and Belief Propagation.

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

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

With $\mathcal{R} = \{R_1, \ldots, R_M\}$ as in Section 2, let $\{\alpha_R(\boldsymbol{x}_R)\}_{R \in \mathcal{R}}$ be a family of nonnegative "local kernels," i.e., $\alpha_R(\boldsymbol{x}_R)$ is a nonnegative real number for each $\boldsymbol{x}_R \in A^R$, and define the global probability density function

$$p(\boldsymbol{x}) = \frac{1}{Z} \prod_{R \in \mathcal{R}} \alpha_R(\boldsymbol{x}_R), \tag{3.1}$$

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

$$Z = \sum_{\boldsymbol{x} \in A^n} \prod_{R \in \mathcal{R}} \alpha_R(\boldsymbol{x}_R). \tag{3.2}$$

The corresponding *probabilistic inference problem* is to compute Z, and the marginal densities

$$p_R(\boldsymbol{x}_R) = \sum_{\boldsymbol{x}_{R^c}} p(\boldsymbol{x}) = \frac{1}{Z} \sum_{\boldsymbol{x}_{R^c}} \prod_{R \in \mathcal{R}} \alpha_R(\boldsymbol{x}_R).$$

for one or more values of R.

If $\mathcal{G} = (V, E, L)$ is a junction graph for $\mathcal{R}$, the *generalized distributive law* [1] is a message-passing algorithm on $\mathcal{G}$ for solving the PI problem, either exactly or approximately. It can be described by its messages and beliefs. If $\{v, u\}$ is an edge of $\mathcal{G}$, a *message* from v to u, denoted by $m_{v,u}(\boldsymbol{x}_{L(v,u)})$, is a nonnegative function on $A^{L(v,u)}$. Similarly if v is a vertex of $\mathcal{G}$, the *belief* at v, denoted by $b_v(\boldsymbol{x}_{L(v)})$, is a probability density on $A^{L(v)}$.

Initially, $m_{v,u}(\boldsymbol{x}_{L(v,u)}) \equiv 1$ for all $\{v, u\} \in E$. The message update rule is

$$m_{v,u}(\boldsymbol{x}_{L(v,u)}) \leftarrow K \sum_{\boldsymbol{x}_{L(v) \setminus L(v,u)}} \alpha_v(\boldsymbol{x}_{L(v)}) \prod_{u' \in N(v) \setminus u} m_{u',v}(\boldsymbol{x}_{L(u',v)}), \tag{3.3}$$

where K is any convenient constant. (In (3.3), $N(v)$ denotes the neighbors of v, i.e., $N(v) = \{u : \{u, v\} \in E\}$.) At any stage, the current beliefs at the vertices v and edges $e = \{u, v\}$ are defined as follows:

$$b_v(\boldsymbol{x}_{L(v)}) = \frac{1}{Z_v} \alpha_v(\boldsymbol{x}_{L(v)}) \prod_{u \in N(v)} m_{u,v}(\boldsymbol{x}_{L(u,v)}). \tag{3.4}$$

$$b_e(\boldsymbol{x}_{L(e)}) = \frac{1}{Z_e} m_{u,v}(\boldsymbol{x}_{L(e)}) m_{v,u}(\boldsymbol{x}_{L(e)}), \tag{3.5}$$

where Z_v and Z_e are the appropriate local normalizing constants, i.e.,

$$Z_v = \sum_{\boldsymbol{x}_{L(v)}} \alpha_v(\boldsymbol{x}_{L(v)}) \prod_{u \in N(v)} m_{u,v}(\boldsymbol{x}_{L(u,v)}) \tag{3.6}$$

$$Z_e = \sum_{\boldsymbol{x}_{L(e)}} m_{u,v}(\boldsymbol{x}_{L(e)}) m_{v,u}(\boldsymbol{x}_{L(e)}). \tag{3.7}$$

The hope is that as the algorithm evolves, the beliefs will converge to the desired marginal probabilities:

$$b_v(\boldsymbol{x}_{L(v)}) \overset{?}{\to} p_{L(v)}(\boldsymbol{x}_{L(v)})$$
$$b_e(\boldsymbol{x}_{L(e)}) \overset{?}{\to} p_{L(e)}(\boldsymbol{x}_{L(e)}).$$

If $\mathcal{G}$ is a tree, the GDL converges as desired in a finite number of steps [1]. Furthermore, a result of Pearl [3] says that if $\mathcal{G}$ is a tree, then the global density $p(\boldsymbol{x})$ defined by (3.1) factors as follows:

$$p(\boldsymbol{x}) = \frac{\prod_{v \in V} p_v(\boldsymbol{x}_{L(v)})}{\prod_{e \in E} p_e(\boldsymbol{x}_{L(e)})}.$$

Comparing this to the definition (3.1), we see that the global normalization constant Z defined in (3.2) can be expressed in terms of the local normalization constants Z_v and Z_e defined in (3.6) and (3.7) as follows:

$$Z = \frac{\prod_v Z_v}{\prod_e Z_e}. \tag{3.8}$$

But what if $\mathcal{G}$ is not a tree? In the remainder of the paper, we shall see that if the GDL converges to a given set of beliefs, these beliefs represent a zero gradient point (conjecturally a local minimum) of a function whose global minimum represents an approximation to the solution to the inference problem. In short, even $\mathcal{G}$ has cycles, the GDL still does something sensible. To see why this is so, we must use some ideas from statistical physics, which we present in the next section.

4. Free Energy and the Boltzmann Distribution.

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*[1] is defined as

$$Z = \sum_{\boldsymbol{x} \in S} e^{-E(\boldsymbol{x})}, \tag{4.1}$$

[1] 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 *free energy* of the system is

$$F = -\ln Z.$$

The free energy is of fundamental importance in statistical physics [8, Chapter 2], and physicists have developed a number of ways for calculating it, either exactly or approximately. We will now briefly describe some of these techniques.

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

$$\widetilde{F}(p) = U(p) - H(p),$$

where $U(p)$ is the average, or internal, energy:

$$U(p) = \sum_{\boldsymbol{x} \in S} p(\boldsymbol{x}) E(\boldsymbol{x}), \tag{4.2}$$

and $H(p)$ is the entropy:

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

We define the *Boltzmann*, or *equilibrium*, distribution as follows:

$$p^B(\boldsymbol{x}) = \frac{1}{Z} e^{-E(\boldsymbol{x})}, \tag{4.3}$$

A routine calculation shows that

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

where $D(p \parallel p^B)$ is the Kullback-Leibler distance between p and p^B. It then follows from [2, Theorem 2.6.3] that

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

with equality if and only if $p(\boldsymbol{x}) = p^B(x)$, which is a classical result from statistical physics [5]. In other words,

$$F = \min_{p(\boldsymbol{x})} \widetilde{F}(p) \tag{4.4}$$

$$p^B(\boldsymbol{x}) = \operatorname*{argmin}_{p(\boldsymbol{x})} \widetilde{F}(p). \tag{4.5}$$

5. The Bethe-Kikuchi Approximation to the Variational Free Energy.

According to (4.4), one method for computing the free energy F is to use calculus to minimize the variational free energy $\widetilde{F}(p)$ over all distributions $p(\boldsymbol{x})$. However, this involves mimimizing a function of the q^n variables $\{p(\boldsymbol{x}) : \boldsymbol{x} \in A^n\}$ subject to the constraint $\sum_{\boldsymbol{x}} p(\boldsymbol{x}) = 1$, which is not attractive, unless q^n is quite small.

Another approach is to *estimate F from above* by minimizing $\widetilde{F}(p)$ over a restricted class of probability distributions. This is the basic idea underlying the *mean field* approach [5; 8, Chapter 4] in which only distributions of the form

$$p(\boldsymbol{x}) = p_1(x_1)p_2(x_2)\cdots p_n(x_n)$$

are considered. The *Bethe-Kikuchi* approximations, which we will now describe, can be thought of as an elaboration on the mean field approach.

In many cases of interest, the energy is determined by relatively short-range interactions, and the Hamiltonian assumes the special form

$$E(\boldsymbol{x}) = \sum_{R \in \mathcal{R}} E_R(\boldsymbol{x}_R), \tag{5.1}$$

where $\mathcal{R}$ is a collection of subsets of $[n]$, as in Sections 2 and 3. If $E(\boldsymbol{x})$ decomposes in this way, the Boltzmann distribution (4.3) factors:

$$p^B(\boldsymbol{x}) = \frac{1}{Z} \prod_{R \in \mathcal{R}} e^{-E_R(\boldsymbol{x}_R)}. \tag{5.2}$$

Similarly, the average energy (cf. (4.2)) can be written as

$$U(p) = \sum_{R \in \mathcal{R}} U(p_R), \tag{5.3}$$

where

$$U(p_R) = \sum_{\boldsymbol{x}_R} p_R(\boldsymbol{x}_R) E_R(\boldsymbol{x}_R). \tag{5.4}$$

Thus the average energy $U(p)$ depends on the global density $p(\boldsymbol{x})$ only through the marginals $\{p_R(\boldsymbol{x}_R)\}$.

One might hope for a similar simplification for $H(p)$:

$$H(p) \overset{?}{\approx} \sum_{R \in \mathcal{R}} H(p_R), \tag{5.5}$$

where

$$H(p_R) = -\sum_{\boldsymbol{x}_R} p_R(\boldsymbol{x}_R) \log p_R(\boldsymbol{x}_R).$$

For example, if

$$\mathcal{R} = \{\{1,2,3\},\{1,4\},\{2,5\},\{3,6\},\{4,5,6\}\}$$

then the hope (5.5) becomes

$$\begin{aligned} &H(X_1,X_2,X_3,X_4,X_5,X_6) \overset{?}{\approx} \\ &H(X_1,X_2,X_3) + H(X_1,X_4) + H(X_2,X_5) + H(X_3,X_6) + H(X_4,X_5,X_6), \end{aligned} \tag{5.6}$$

but this is clearly false, if only because the random variables X_i occur unequally on the two sides of the equation. This is where the junction graph comes in. If, instead of (5.5) we substitute the *Bethe-Kikuchi approximation* to $H(p)$ (relative to the junction graph $\mathcal{G} = (V, E, L)$):

$$H(p) \approx \sum_{v \in V} H(p_v) - \sum_{e \in E} H(p_e), \tag{5.7}$$

where for simplicity we write p_v instead of $p_{L(v)}$, etc., the junction graph condition guarantees that each X_i is counted just once. (In a tree, the number of vertices is exactly one more than the number of edges.) For example, using the junction graph of Figure 1, (5.7) becomes

$$\begin{aligned} H(X_1, X_2, X_3, X_4, X_5, X_6) &\overset{?}{\approx} \\ &H(X_1, X_2, X_3) + H(X_1, X_4) + H(X_2, X_5) + H(X_3, X_6) + H(X_4, X_5, X_6) \\ &- H(X_1) - H(X_2) - H(X_3) - H(X_4) - H(X_5) - H(X_6), \end{aligned} \tag{5.8}$$

which is more plausible (though it needs to be investigated by an information theorist).

In any case, the *Bethe-Kikuchi approximation* with respect to the junction graph $\mathcal{G} = (V, E, L)$ to the variational free energy $\widetilde{F}(p)$ is defined as (cf. (5.3) and (5.7))

$$\widetilde{F}_{BK}(p) = \sum_{v \in V} U_{L(v)}(p_v) - \left(\sum_{v \in V} H(p_v) - \sum_{e \in E} H(p_e) \right). \tag{5.9}$$

The important thing about the BK approximation $\widetilde{F}_{BK}(p)$ in (5.9) is that it depends only on the marginal probabilities $\{p_v(\boldsymbol{x}_{L(v)})\}$, $\{p_e(\boldsymbol{x}_{L(e)})\}$, and not on the full global distribution $p(\boldsymbol{x})$. To remind ourselves of this fact, we will write $\widetilde{F}_{BK}(\{p_v, p_e\})$ instead of $\widetilde{F}_{BK}(p)$.

One plausible way to estimate the free energy is thus (cf. (4.4)) $F \approx F_{BK}$, where

$$F_{BK} = \min_{\{b_v, b_e\}} \widetilde{F}_{BK}(\{b_v, b_e\}), \tag{5.10}$$

where the b_v's and the b_e's are trial marginal probabilites, or "beliefs." The BK "approximate beliefs" are then the optimizing marginals:

$$\{b_v^{BK}, b_e^{BK}\} = \underset{\{b_v, b_e\}}{\operatorname{argmin}} \widetilde{F}_{BK}(\{b_v, b_e\}). \tag{5.11}$$

Computing the minimum in (5.10) is still not easy, but we can begin by setting up a Lagrangian:

$$\begin{aligned} \mathcal{L} = &\sum_{v \in V} \sum_{\boldsymbol{x}_{L(v)}} b_v(\boldsymbol{x}_{L(v)}) E_{L(v)}(\boldsymbol{x}_{L(v)}) \\ &+ \sum_{v \in V} \sum_{\boldsymbol{x}_{L(v)}} b_v(\boldsymbol{x}_{L(v)}) \log b_v(\boldsymbol{x}_{L(v)}) \\ &- \sum_{e \in E} \sum_{\boldsymbol{x}_{L(e)}} b_e(\boldsymbol{x}_{L(e)}) \log b_e(\boldsymbol{x}_{L(e)}) \\ &+ \sum_{(u,v) \in E} \sum_{\boldsymbol{x}_{L(u,v)}} \lambda_{u,v}(\boldsymbol{x}_{L(u,v)}) \left(\sum_{\boldsymbol{x}_{L(v) \setminus L(u,v)}} b_v(\boldsymbol{x}_{L(v)}) - b_e(\boldsymbol{x}_{L(u,v)}) \right) \\ &+ \sum_{v \in V} \mu_v \left(\sum_{\boldsymbol{x}_{L(v)}} b_v(\boldsymbol{x}_{L(v)}) - 1 \right) \\ &+ \sum_{e \in E} \mu_e \left(\sum_{\boldsymbol{x}_{L(e)}} b_e(\boldsymbol{x}_{L(e)}) - 1 \right). \end{aligned}$$

Here the Lagrange multiplier $\lambda_{u,v}(\boldsymbol{x}_{L(u,v)})$ enforces the constraint

$$\sum_{\boldsymbol{x}_{L(v)\setminus L(u,v)}} b_v(\boldsymbol{x}_{L(v)}) = b_e(\boldsymbol{x}_{L(u,v)}),$$

the Lagrange multiplier μ_v enforces the constraint

$$\sum_{\boldsymbol{x}_{L(v)}} b_v(\boldsymbol{x}_{L(v)}) = 1,$$

and the Lagrange multiplier μ_e enforces the constraint

$$\sum_{\boldsymbol{x}_{L(e)}} b_e(\boldsymbol{x}_{L(e)}) = 1.$$

Setting the partial derivative of $\mathcal{L}$ with respect to the variable $b_v(\boldsymbol{x}_{L(v)})$ equal to zero we obtain, after a little rearranging,

$$\log b_v(\boldsymbol{x}_{L(v)}) = k_v - E_{L(v)}(\boldsymbol{x}_{L(v)}) - \sum_{u\in N(v)} \lambda_{v,u}(\boldsymbol{x}_{L(u,v)}), \tag{5.12}$$

(where $k_v = -1 - \mu_v$) for all vertices v and all choices for $\boldsymbol{x}_{L(v)}$. Similarly, setting the partial derivative of $\mathcal{L}$ with respect to the variable $b_e(\boldsymbol{x}_{L(e)})$ equal to zero we obtain

$$\log b_e(\boldsymbol{x}_{L(e)}) = k_e - \lambda_{v,u}(\boldsymbol{x}_{L(e)}) - \lambda_{u,v}(\boldsymbol{x}_{L(e)}), \tag{5.13}$$

(where $k_e = -1 + \mu_e$) for all edges $e = \{u, v\}$ and all choices for $\boldsymbol{x}_{L(e)}$. The conditions (5.12) and (5.13) are necessary conditions for the attainment of the minimum in (5.10). A set of beliefs and Lagrange multipliers satisfying these conditions may correspond to a local minimum, a local maximum, or neither. In every case, however, we have a *stationary point* for $\widetilde{F}_{BK}$.

Given a stationary point of $\widetilde{F}_{BK}$, if we define local kernels $\alpha_{L(v)}(\boldsymbol{x}_{L(v)})$ and messages $m_{u,v}(\boldsymbol{x}_{L(u,v)})$ as follows:

$$E_v(\boldsymbol{x}_{L(v)}) = -\ln \alpha_{L(v)}(\boldsymbol{x}_{L(v)}) \tag{5.14}$$

$$\lambda_{v,u}(\boldsymbol{x}_{L(v,u)}) = -\ln m_{u,v}(\boldsymbol{x}_{L(u,v)}), \tag{5.15}$$

the stationarity conditions (5.12) and (5.13) then become

$$b_v(\boldsymbol{x}_{L(v)}) = K_v \alpha_{L(v)}(\boldsymbol{x}_{L(v)}) \prod_{u\in N(v)} m_{u,v}(\boldsymbol{x}_{L(u,v)}) \tag{5.16}$$

$$b_e(\boldsymbol{x}_{L(e)}) = K_e m_{u,v}(\boldsymbol{x}_{L(u,v)}) m_{v,u}(\boldsymbol{x}_{L(v,u)}), \tag{5.17}$$

which arre the same as the GDL belief rules (3.4) and (3.5). (The constants $K_v = e^{k_v}$ and $K_e = e^{k_e}$ are uniquely determined by the conditions $\sum_{\boldsymbol{x}_{L(v)}} b_v(\boldsymbol{x}_{L(v)}) = 1$ and $\sum_{\boldsymbol{x}_{L(e)}} b_e(\boldsymbol{x}_{L(e)}) = 1$.)

Next, if we choose $u \in N(v)$, and sum (5.16) over all $\boldsymbol{x}_{L(v)\setminus L(v,u)}$, thereby obtaining an expression for $b_e(\boldsymbol{x}_{L(e)})$, where $e = \{u, v\}$, and then set the result equal to the right side of (5.17), we obtain, after a short calculation,

$$m_{v,u}(\boldsymbol{x}_{L(u,v)}) = \frac{K_v}{K_e} \sum_{\boldsymbol{x}_{L(v)\setminus L(v,u)}} \alpha_{L(v)}(\boldsymbol{x}_{L(v)}) \prod_{u'\in N(v)\setminus u} m_{u',v}(\boldsymbol{x}_{L(u,v)}), \tag{5.18}$$

which is the same as the GDL message update rule (3.3). In other words, we have proved

Theorem 2. *If a set of beliefs $\{b_v, b_e\}$ and Lagrange multipliers $\{\lambda_{u,v}\}$ is a stationary point for the BK free energy defined on $\mathcal{G}$ with energy function given by (5.1), then these same beliefs, together with the set of messages defined by (5.15), are a fixed point for the GDL of $\mathcal{G}$ defined by the local kernels defined by (5.14). Conversely, if a set of beliefs $\{b_v, b_e\}$ and messages $\{m_{u,v}\}$ is a fixed point of the GDL on a junction graph $\mathcal{G}$ with local kernels $\{\alpha_R(\boldsymbol{x}_R)\}$, then these same beliefs, together with the Lagrange multipliers $\{\lambda_{u,v}\}$ defined by (5.15) are a stationary point for the BK free energy defined on $\mathcal{G}$ with energy function defined by (5.14).*

Finally we state the following two theorems without proof.

Theorem 3. *If $|V| \geq |E|$ (in particular, if $\mathcal{G}$ is a tree or if $\mathcal{G}$ has only one cycle), then $\widetilde{F}_{BK}$ is convex $\cup$, and hence has only one stationary point, which is a global minimum.*

Theorem 4. *The value of $\widetilde{F}_{BK}$ corresponding to a stationary point depends only on the Lagrange multipliers $\{k_v, k_e\}$, as follows:*

$$\widetilde{F}_{BK} = \sum_v k_v - \sum_e k_e.$$

(This result should be compared to (3.8).)

Acknowledgement.

The authors thank Dr. J. C. MacKay for his advice on the material in Section 4.

References.

1. S. M. Aji and R. J. McEliece, "The generalized distributive law," *IEEE Trans. Inform. Theory*, vol. 46, no. 2 (March 2000), pp. 325–343.

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

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

4. R. P. Stanley, *Enumerative Combinatorics, vols. I and II.* (Cambridge Studies in Advanced Mathematics 49, 62) Cambridge: Cambridge University Press, 1997–98.

5. 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.

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

7. 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/`

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