Minimal Graphical Representation of Kikuchi Regions*

Payam Pakzad  
EECS Department  
University of California,  
Berkeley, CA 94720  
payamp@eecs.berkeley.edu

Venkat Anantharam  
EECS Department  
University of California,  
Berkeley, CA 94720  
ananth@eecs.berkeley.edu

Abstract

It was shown recently that the well known belief propagation algorithms for posterior probability evaluation can be viewed as algorithms that aim to minimize certain approximations to the variational free energy in a statistical physics context. Specifically, the fixed points of belief propagation algorithms are shown to coincide with the stationary points of Bethe's approximate free energy subject to certain consistency constraints. Bethe's approximation is known to be a special case of a more general class of approximations called Kikuchi free energy approximations. A more general class of belief propagation algorithms was thus introduced which corresponds to algorithms that aim to minimize a general Kikuchi approximate free energy.

In this paper we first review this circle of ideas. Specifically, given an arbitrary collection of regions, i.e. proper subsets of a set of state variables, and a collection of functions of the configuration of state variables over these regions, we define a general constrained minimization problem corresponding to the general Kikuchi approximation whose stationary points approximate marginals over these regions of the product function, and we specify a general class of local message-passing algorithms along the edges of a graphical representation of the collection of Kikuchi regions, which attempt to solve that problem. Our main contribution then follows, which is to introduce a suitable minimal graphical representation of the collection of regions. Iterative message-passing algorithms on the graph we construct involve fewest message updates at each iteration. We also prove that exactness of Kikuchi approximation of marginals depends directly on this graph being cycle-free.

1 Background

Let $x := (x_1, \ldots, x_N)$ be a collection of state variables where $x_i$ takes values in $\{0, \ldots, q_i - 1\}$ respectively, with $q_i \geq 2$; In a statistical physics context $x_i$ is interpreted as the 'spin' of the particle at position $i$ in a system of $N$ particles. Let $b(x)$ denote a probability distribution on configurations of states and let $\xi_x$ be a real function of $x$, interpreted as the energy of the system in configuration $x$.

*This work was supported by grants from (ONR/MURI) N00014-1-0637, (NSF) SBR-9873086, (DARPA) F30602-00-2-0538, California Micro Program, Texas Instruments, Marvell Technologies and ST MicroElectronics.
In thermal physics one defines the *variational free energy* as the following functional of the distribution:

\[ F(b(x)) := U(b(x)) - S(b(x)) \tag{1} \]

where \( U := \sum_x b(x)\varepsilon_x \) is the average energy and \( S := -\sum_x b(x)\ln(b(x)) \) is the entropy of the system.

It can be shown that the free energy \( F \) is uniquely minimized when \( b(x) \) equals the Boltzmann distribution

\[ B(x) := \frac{e^{-\varepsilon_x}}{Z} \tag{2} \]

Here \( Z \) is a normalizing factor and is called the *partition function*. Then we have

\[ F_0 := \min_{b(x)} F(b(x)) = F(B(x)) = -\ln(Z) \tag{3} \]

Equation (3) is of special interest. Physicists are interested in finding \( F_0 \) (as a function of a temperature variable which we have omitted here, but appears as a scaling factor on the energy) since thermodynamical properties of the system can be derived from it. For the subsequent discussion, the main point to take away is that equation (3) does not generally prescribe a practical way to compute \( F_0 \) as it involves minimization over the exponentially large domain of distributions \( b(x) \).

In statistics, coding, artificial intelligence, and estimation theory, it turns out that posterior probability calculation, when suitably viewed, reduces to finding the marginals of the Boltzmann distribution \( B(x) \) with respect to certain subsets of the state variables. This connection is developed in more detail in Section 3. Let \( R \) be a collection of *regions*, i.e. proper subsets of \( \{1, \cdots, N\} \), and let \( \Delta_R \) denote the collection of probability distributions over the configurations restricted to the regions which are the true marginals of a complete distribution, i.e. a collection \( \{b_r(x_r), r \in R\} \) belongs to \( \Delta_R \) if and only if there exists a distribution \( b(x) \) s.t. \( \forall r \in R, b_r(x_r) = \sum_{x \setminus x_r} b(x) \). Then, if we were primarily interested in the marginals of the Boltzmann distribution with respect to the regions, we could rewrite (3) as

\[ F_0 = \min_{\{b_r(x_r)\} \in \Delta_R} F_R(\{b_r(x_r)\}) \]

\[ \{B_r(x_r)\} = \arg \min_{\{b_r(x_r)\} \in \Delta_R} F_R(\{b_r(x_r)\}) \tag{4} \]

Here \( F_R(\{b_r(x_r)\}) \) should itself be viewed as a minimum of \( F(b(x)) \) from (1) over the set of distributions \( b(x) \) having marginals \( \{b_r(x_r), r \in R\} \). The minimizers \( \{B_r(x_r), r \in R\} \) in the above equation are of course the marginals of the Boltzmann distribution.

## 2 Decomposable Models and Kikuchi Approximation

In the applications involving posterior probability calculation, the distribution to be marginalized is often given in terms of a product of certain prior distributions and various conditional probability distributions, both of which depend only on certain subsets of the variables. This also corresponds, as will become clearer in Section 3, to the important class of statistical physics problems where it is given a priori that the energy function \( \varepsilon_x \) can be decomposed as

\[ \varepsilon_x = \sum_{r \in R} E_r(x_r) \tag{5} \]
for some set of functions \( \{ E_r(x_r), r \in R \} \). Some familiarity with the Hammersley-Clifford theorem for Markov random fields, (see e.g. Theorem 1.1 on pg. 7 of [7],) will help develop an appreciation for the importance of this case, but is not necessary for the rest of the paper. Examining the variational free energy of (1) in an attempt to express it as a functional of the collection of marginals \( \{ b_r(x_r), r \in R \} \) of the complete distribution, – as we half-heartedly did in (4), – we see that in this case the average energy decomposes nicely as

\[
U(b) = \sum_{r \in R} \sum_{x_r} b_r(x_r) E_r(x_r).
\]

In general however, the entropy term of the free energy (1) cannot be decomposed in this form. The idea of the Kikuchi approximation is to replace the entropy term by an approximation of the form

\[
S(b) \approx \sum_{r \in R} c_r S_r(b_r)
\]

where \( c_r \)'s are suitable constant factors, and \( S_r(b_r) := -\sum_{x_r} b_r(x_r) \log(b_r(x_r)) \) is the regional entropy associated with a region \( r \in R \).

View \( R \cup \{1, \cdots, N\} \) as a poset with partial ordering of inclusion. The specific choice of entropy approximation used in the Kikuchi approximation invokes the Möbius inversion formula (see [8]) on this poset and sets\( c_r = -\mu(r, \{1, \cdots, N\}) \) where \( \mu \) is the Möbius function. The rationale behind this is developed in [4] following [3] and will not be repeated here. Note that one has

\[
c_r = 1 - \sum_{s \in A(r)} c_s
\]

where \( A(r) := \{ s \in R : r \subset s \} \) is the set of ancestors of \( r \). Following [9] we may also call \( c_r \)'s the overcounting factors.

Kikuchi’s approximate free energy uses the above approximation of entropy together with average energy form of (6) in (1) (cf. equation (35) in [9]) to write:

\[
F^K_R(\{b_r\}) := \sum_{r \in R} \sum_{x_r} (b_r(x_r) E_r(x_r) + c_r b_r(x_r) \log(b_r(x_r)))
\]

Note however that the region over which we wish to minimize this approximate function is defined in terms of the complicated requirement that it be comprised of the set of marginals of a probability distribution over configurations. Thus a second step in the Kikuchi approximation method is to approximate the constraint set \( \Delta_R \) of equation (4).

Note that the marginals of a distribution function will be consistent with each other and will normalize to 1. Therefore a reasonable choice for an approximate constraint set is the following:

\[
\Delta^K_R := \{ (b_r(x_r), r \in R) : \forall t, u \in R \text{ s.t. } t \subset u, \sum_{x_u \setminus t} b_u(x_u) = b_t(x_t) \text{ and } \sum_{x_t} b_t(x_t) = 1 \}
\]

Note that in general the constraints of \( \Delta^K_R \) are not enough to guarantee that every collection of pseudo-marginals \( \{ b_r, r \in R \} \in \Delta^K_R \) is in fact the collection of the marginals of a single distribution function \( b(x) \) : it is not hard to construct collections of pseudo-marginals that satisfy all the consistency constraints of (10) but are nevertheless not
the marginals of any distribution. In this connection, we might note that as part of our discussion, in Section 6, we discuss conditions on $R$ that guarantee that the Kikuchi functional $F^K_R$ equals the free energy $F(b)$ and that the constraint set $\Delta^K_R$ equals the collection of marginals of a single distribution function.

Using approximations (9) and (10) we have the following optimization problem:

$$F_0 \simeq \min_{\{b_r(x_r)\} \in \Delta^K_R} F^K_R (\{b_r(x_r)\})$$

with \(\{b^*_r(x_r)\} = \arg \min_{\{b_r(x_r)\} \in \Delta^K_R} F^K_R (\{b_r(x_r)\})\) \hspace{1cm} (11)

One might be interested only in the value of the Kikuchi approximate free energy as in statistical physics or one might be interested in both this and a collection of pseudo-marginals where this minimum is achieved, as is typical in problems in estimation. One way to phrase the key question in the latter context is to ask how close the $b^*_r$'s are to the marginals $B_r$ of the Boltzmann distribution.

3 A General Class of Constrained Minimization Problems

In this section we explore how the Kikuchi approximation method may be massaged to provide good approaches to posterior probability calculations. In the process we also elucidate the connection between the statistical physics formulation and the viewpoint of estimation theory as was promised earlier.

Let $R_0$ be a collection of regions, and \(\{E_r^0(x_r), r \in R_0\} \) be a collection of functions so that, as in (5), $\varepsilon_x = \sum_{r \in R_0} E_r^0(x_r)$. Let $R$ be another collection of regions so that $\forall r \in R_0, \exists r' \in R \text{ s.t. } r \subseteq r'$. Then one can always form a collection of functions \(\{E_r(x_r), r \in R_0\} \) so that equation (5) holds.\(^1\)

Now for each $r \in R$, define the potentials $\alpha_r(x_r) := e^{-E_r(x_r)}$, and $\beta_r(x_r) := \prod_{s \subseteq r} \alpha_r(x_r)$. Then the Boltzmann distribution takes the form of a product function of the potentials:

$$B(x) = \prod_{r \in R} \frac{e^{-E_r(x_r)}}{Z} = \prod_{r \in R} \frac{\alpha_r(x_r)}{Z} = \prod_{r \in R} \frac{\beta_r(x_r)^{c_r}}{Z}$$ \hspace{1cm} (12)

where the last equality follows from the fact that $\sum_{r \in R, s \subseteq r} c_r = 1$ for all $s \in R$.

In an estimation theory context one is typically given some prior distributions and certain conditional distributions, corresponding to the collection $\{E_r^0(x_r), r \in R_0\}$. One then has the freedom to choose $R$, as long as it includes the regions over which the desired posterior probability distributions are defined, and so that $\forall r \in R_0, \exists r' \in R \text{ s.t. } r \subseteq r'$.\(^2\)

One then also has flexibility in choosing a collection of functions $\{E_r(x_r), r \in R_0\}$ so that equation (5) holds. These choices then specify the Kikuchi approximation, both in terms of the approximation to the variational free energy (9), and in terms of the approximation to the constraint set (10). It is also evident that (11) as an approximation method can be applied for any given $F^K_R$ and $\Delta^K_R$; better choices are defined by the fact that they result in better approximations.

\(^1\)One way to do this is to define $E_r(x_r) := \sum_{s \in R, E_s^0(x_s)}$ for each maximal $r \in R$, and $E_t(x_t) := 0$ for non-maximal elements $t \in R$. The way this assignment is done, however, can impact the quality of the approximations to (4) provided by (11).
Once these choices have been made we are concerned with an example of a general class of constrained minimization problems as above, which are specified by a poset \( R \) of regions, and local potential functions \( \alpha_r(x_r) \) for each \( r \in R \). It is natural to represent poset \( R \) with its Hasse diagram \( G_R \) (see \cite{8}). This is a directed acyclic graph (DAG), whose vertices are the elements of \( R \) and whenever \( t \) covers \( u \) in \( R \), there is an edge \( (t \to u) \) pointing from \( t \) to \( u \). We now associate each edge \( (t \to u) \) of \( G_R \) with a local consistency constraints \( \sum_{x_t} b_t(x_t) = b_u(x_u) \). We refer to this constraint as the edge-constraint of \( (t \to u) \). Then the collection of edge-constraints of \( G_R \) is a sufficient representation of \( \Delta^K_R \), i.e. \( \{b_r, r \in R\} \in \Delta^K_R \) if \( \sum_{x_t} b_t(x_t) = b_u(x_u) \) for each edge \( (t \to u) \) of \( G_R \). As we shall see in the next section, there exist local message-passing algorithms along the edges of \( G_R \), whose fixed points are the minimizers \( \{b_r^*\} \) of the constrained minimization problem (11).

Since the problem formulation was motivated in terms of approximating the collection of marginals of a Boltzmann distribution, it is important to specify which choices yield good approximations of the marginals. In the rest of this section we explore this issue. The remarks we make do not directly impact the rest of the paper, where it is assumed that some such choice has already been made.

First of all, to preserve the low complexity of minimization problem (11), one may restrict attention to collections of regions \( R \) that have the same maximal regions as \( R_0 \).

Secondly, it certainly seems that minimization with more local consistency constraints on \( \{b_r(x_r)\} \) should result in better approximations, since true marginals would satisfy all such constraints. Therefore one might conclude that for a given collection of maximal regions of \( R_0 \), augmenting them by introducing additional subregions to form \( R \), where the \( E_r \)'s corresponding to the augmented subregions are taken to be zero, should improve the approximation (at the expense of slightly increasing the complexity of \( G_R \) and its corresponding algorithm).

Thirdly, as we discussed in \cite{4}, it is natural to require that \( R \) be connected at the level of subsets of size \( n \) or less for some integer \( n \geq 1 \), i.e. for each \( s \subset \{1, \cdots, N\} \) with \( |s| \leq n \), all the regions containing \( s \) be connected in the Hasse diagram (Property (An)). This will ensure that the beliefs \( b_r(x_r) \) at all the regions \( r \) which contain \( s \) will be consistent at the level of variables \( x_s \).

On the other hand one might also insist that acceptable approximations of the entropy term (7) are those that are balanced for subsets of size \( n \) or less for some integer \( n \), in the sense that each subset \( s \subset \{1, \cdots, N\} \) with \( |s| \leq n \) appears the same number of times on the two sides of the equality sign of (7):

\[
\sum_{r: s \subseteq r} c_r = 1 \quad \text{for each } s \subset \{1, \cdots, N\}, |s| \leq n \quad \text{(Property (Bn))}
\]

These conditions are expected to give progressively better approximate solutions. It is noteworthy that the original cluster variational method of Kikuchi (see \cite{3} and \cite{9}) involves a collection of regions that is closed under intersection; it can be shown that any collection of regions which is closed under intersection satisfies (An) and (Bn) for all \( n \).

Finally, the special case when the Hasse diagram \( G_R \) has depth 2, i.e. there are no distinct \( r, s, t \in R \) such that \( r \subset s \subset t \), is called the general Bethe case. In this case \( G_R \) can be thought of as a hypergraph of which Belief Propagation algorithm is defined.

---

\footnote{We say \( t \) covers \( u \) in \( R \) and write \( u \prec t \), if \( u, t \in R \), \( u \subset t \) and \( \bar{b}v \in R \) s.t. \( u \subset v \subset t \).}

\footnote{Expansion of the maximal regions corresponds to ‘clustering’ methods discussed in \cite{6}, introducing a new dimension to the problem of trading off complexity for accuracy.}
4 Lagrange Multipliers and Iterative Solutions

Lagrange’s method can be used to solve the constrained minimization problem (11). We form the Lagrangian:

\[
\mathcal{L} := \sum_{r \in R} \sum_{x_r} (-b_r(x_r) \ln(\alpha_r(x_r)) + c_r b_r(x_r) \ln(b_r(x_r)))
+ \sum_{r \in R} \sum_{t < r} \lambda_{rt}(x_t)(b_t(x_t) - \sum_{x_{r \cap t}} b_r(x_r)) + \sum_{r \in R} \kappa_r \left( \sum_{x_r} b_r(x_r) - 1 \right)
\]  

(13)

where coefficients \( \lambda_{rt}(s_t) \) enforce consistency constraints, and coefficients \( \kappa_r \) enforce normalization constraints, and as before \( t \prec r \) means that \( r \) covers \( t \). Note that since the edge-constraints of \( G_R \) are a sufficient representation of \( \Delta^K_R \) as discussed before, we need only define \( \lambda_{rt} \) for pairs \( r, t \in R \) with \( t \prec r \), i.e. along the edges of \( G_R \).

Setting partial derivative \( \partial \mathcal{L} / \partial b_r(x_r) = 0 \) for each \( r \in R \) gives an equation for \( b_r(x_r) \) in terms of \( \lambda_{ur} \)'s and \( \lambda_{rt} \)'s. The consistency constraints give update rules for each \( \lambda_{rt} \) in terms of other \( \lambda \) multipliers. Once a set of messages \( m_{rt} \) (from \( r \) to \( t \), for each edge \( (r \rightarrow t) \) of \( G_R \)) has been defined in terms of the Lagrange multipliers \( \lambda_{rt} \)'s, these update rules define an iterative algorithm whose fixed points are the stationary points of the given constrained minimization problem.

One particularly nice such algorithm is the ‘Parent-to-Child’ algorithm discussed in [10] which defines the messages so that belief \( b_r(x_r) \) depends only on the outside messages to a subregion of \( r \):

\[
b_r(x_r) = k \beta_r(x_r) \left( \prod_{p : r < p} m_{pr}(x_r) \right) \left( \prod_{d : d \in r} \prod_{p' : d < p'} \prod_{p' \in r} m_{p'd}(x_d) \right)
\]  

(14)

In particular, for each edge \((p \rightarrow r)\) of \( G_R \) the message \( m_{pr}(x_r) \) is defined as \( m_{pr}(x_r) := e^{-\mu_{pr}(x_r)} \), where \( \{\mu_{pr}(x_r)\} \) is a ‘rotated’ version of the original Lagrange multipliers \( \{\lambda_{pr}(x_r)\} \) (see [10], [5] for detailed derivation.)

The update rule for message \( m_{pr}(x_r) \) is obtained from consistency constraint \( b_r(x_r) = \sum_{x_{r \setminus r \cap p}} b_p(x_p) \), where \( b_r \) and \( b_p \) are expressed in terms of messages using equation (14).

The belief propagation algorithm of [6] can be seen as the restriction of the above algorithm in the Bethe case (see [9] and [2]).

5 Uniqueness of Solution

In this section we recall the results regarding the uniqueness of solutions to the optimization problem (11), which we reported in [4].

The Kikuchi free energy (9) constrained on \( \{b_r\} \in \Delta^R_K \) is bounded below and hence the constrained minimization problem (11) always has a global minimum. Therefore, as discussed in Section 4, the message passing algorithms derived from Lagrangian (13) always possess at least one fixed point (see [11] for an algorithm that is guaranteed to find a minimum of \( F_K \)).

The following result gives sufficient conditions on \( R \) for the problem (11) to have precisely one minimum:
Theorem 1. The Kikuchi free energy functional (9) is convex on $\Delta^K_R$ (and hence the constrained minimization problem has a unique solution) if the overcounting factors $c_r, r \in R$ satisfy:

$$\forall S \subset R, \sum_{r \in R \mid \exists s \in S \cup r} c_r \geq 0$$ \hspace{1cm} (15)

In words, for any subset $S$ of $R$, the sum of overcounting factors of elements of $S$ and all their ancestors in $R$ must be nonnegative.

Remember that for maximal regions $r \in R$, $c_r = 1$, and in the Bethe case, for the non-maximal regions $t \in R$, $c_t = 1 - (\#$ of parents of $t)$. Thus we have

Corollary 2. (cf. Theorem 3 in [2]) In the Bethe case, the constrained minimization problem (11) has a unique solution if the graphical representation $G_R$ of $R$ has at most one loop.

6 Graphical Representation of the Collection of Regions

The results of [4] which we recalled in the previous section refer to the uniqueness of solution of the constrained minimization problem (11). However, one is further interested in the conditions under which these solutions are the exact marginals of the product function (12).

In this section we define a minimal graphical representation of a given collection $R$ of regions, and show that exactness of approximations obtained from (11) corresponds to existence of loops in this graph. In fact, we will show that in the loop-free case, this graph is a junction tree and so the message-passing algorithms of type discussed in Section 4 correspond to (a variation of) the junction tree algorithm.

As before, let $R$ be a poset of regions with partial ordering of inclusion.

For each node $r \in R$, define:

Ancestors: $A(r) := \{s \in R : r \subset s\}$

Descendents $D(r) := \{s \in R : s \subset r\}$

Parents $P(r) := \{s \in R : r \prec s\}$

Children $C(r) := \{s \in R : s \prec r\}$

Family $F(r) := \{r\} \cup A(r)$

Also define a depth function for each region $r \in R$ as:

$$d(r) := \begin{cases} 
0 & \text{if } r \text{ is maximal in } R \\
1 + \max_{s \in P(r)} d(s) & \text{otherwise}
\end{cases}$$

Similarly we define the depth of each edge $(t \to u)$ of the Hasse diagram $G_R$, as the depth of the child node $u$: $d((t \to u)) := d(u)$.

For a graph $G$, denote by $E(G)$ the set of edges of $G$.

As mentioned before, Hasse diagram $G_R$ is the most natural graphical representation of poset $R$, with the property that the collection of edge-constraints of $G_R$ is a sufficient representation of $\Delta^K_R$. 

1591
Hasse diagram uses the transitivity of partial ordering to represent a poset in the most compact form: a relation \( u \subseteq t \) exists if and only if there is a directed path from \( t \) to \( u \), and further, removal of any of the edges results in some of the relations not being represented. Our local consistency constraints also have the transitivity property, i.e. if \( \sum_{x_i \setminus a} b_t(x_i) = b_u(x_u) \) and \( \sum_{x_i \setminus v} b_u(x_u) = b_v(x_v) \) then \( \sum_{x_i \setminus a} b_t(x_i) = b_v(x_v) \), which is why the edge-constraints of \( G_R \) are sufficient to represent \( \Delta_R^K \). On the other hand, local consistency relations satisfy a property other than transitivity which can be used to further reduce the representation of \( \Delta_R^K \). In particular, for \( r, s, t, u \in R \) s.t. \( u \subseteq s \subseteq r \) and \( u \subseteq t \subseteq r \), if \( \sum_{x_i \setminus a} b_t(x_i) = b_s(x_s) \), \( \sum_{x_i \setminus a} b_t(x_i) = b_u(x_u) \) and \( \sum_{x_i \setminus a} b_t(x_i) = b_t(x_t) \) then \( \sum_{x_i \setminus a} b_t(x_i) = b_u(x_u) \) (Property \( \Diamond \)), so that the last edge-constraint can be removed from the graph. We make this precise as follows:

**Definition 1.** Edges \((u \rightarrow r)\) and \((v \rightarrow r)\) are defined to be equivalent for removal, and denoted \((u \sim r) \sim (v \rightarrow r)\) if there exists a sequence \((t_0 \rightarrow r), \cdots, (t_k \rightarrow r)\) of edges in \( G_R \), with \( t_0 = u \) and \( t_k = v \) and with the property that \( \forall i = 1, \cdots, k, \mathcal{A}(t_{i-1}) \cap \mathcal{A}(t_i) \neq \emptyset \), i.e. \( \exists w_i \in R \) s.t. \( t_{i-1} \subseteq w_i \) and \( t_i \subseteq w_i \).

Then it is easy to verify that this relation \( \sim \) is indeed an equivalence relation, and hence, for each region \( r \in R \), the collection of all the edges leading to \( r \) can be partitioned into equivalence classes (of edges that are equivalent for removal).

Now from each such equivalence class \( \{(t_1 \rightarrow r), \cdots, (t_m \rightarrow r)\} \), remove all but one (representative) edge from the Hasse diagram \( G_R \). Denote the resulting graph by \( S_R \).

Note that graph \( S_R \) is not unique, since the representative edge of each equivalence class can be arbitrarily chosen. However, the number of the edges of any choice of \( S_R \) is unique and equals the total number of equivalence classes of edges for removal. Further, the number of loops of any instance of \( S_R \) is the same. As we will see shortly, all choices of \( S_R \) result in equivalent, minimal graphical representations of \( R \). All results in this section apply to every choice of \( S_R \). See [5] for proofs of results in this section.

**Lemma 3.** If \( r, t \in R \) and \( r \subseteq t \), then there is a path in \( S_R \) between \( r \) and \( t \) consisting only of nodes that contain \( r \).

**Proposition 4.** Edge-constraints of \( S_R \) are a minimal representation of the constraint set \( \Delta_R^K \), i.e. a collection of pseudo-marginals \( \{b_r, r \in R\} \) lies in \( \Delta_R^K \) if and only if it satisfies all the edge-constraints of \( S_R \), and further, removal of any of the edges of \( S_R \) results in misrepresentation of \( \Delta_R^K \).

As we have seen, to solve the constraint minimization problem one forms the Lagrangian, introducing multipliers \( \lambda_r(x_r) \) for each edge \((t \rightarrow r)\) of \( S_R \). Since \( S_R \) has fewer edges than \( G_R \), algorithms based on \( S_R \) require fewer message updates for each iteration than those based on \( G_R \).

**Definition 2.** Let \( \{r_1, \cdots, r_M\} \) be a collection of subsets of the index set \( \{1, \cdots, N\} \). A tree/forest with vertices \( \{r_1, \cdots, r_M\} \) is called a junction tree/forest if the subgraph consisting of all the vertices that contain an index \( i \in \{1, \cdots, n\} \) is connected.

Although junction trees are traditionally defined as undirected trees, in the above definition we do not make distinction between directed and undirected graphs; we call a directed graph a junction tree if replacing all the directed edges with undirected ones yields a junction tree in the usual sense.
A well-known result indicates that Belief Propagation algorithm converges to the exact marginals\textemdash in finite time\textemdash if the \textquote{underlying graph} is a junction tree (See e.g. [6], [1]). It turns out that the same can be said about the exactness of the message-passing algorithms of the type discussed in Section 4 on Hasse diagram $G_R$, but this is not a very strong result, as very rarely $G_R$ will be loop-free. In fact many collections of regions that can be put on a junction tree result in Hasse diagrams that have loops. For example $R = \{\{123\}, \{234\}, \{345\}, \{23\}, \{34\}, \{3\}\}$ will have a loop in the Hasse diagram, but can be easily handled as a junction tree.

It turns out that not all the loops of $G_R$ are \textquote{bad} loops that cause trouble for the message-passing algorithm. In fact these \textquote{bad} loops are precisely the loops that cannot be broken when one creates $S_R$. The following results make this precise:

Let $\{r_1, \cdots, r_M\}$ be maximal elements of $R$. For the rest of this section we assume that $R$ includes $r_i \cap r_j$ for each $i, j \in \{1, \cdots, M\}$, so $R$ has property (A1) of Section 3. Then

**Proposition 5.** If $S_R$ has no loops, then $S_R$ is a junction forest and hence $\{r_1, \cdots, r_M\}$ can be put on a junction tree.

Interestingly, the converse to the above proposition is also true:

**Proposition 6.** If the maximal elements $\{r_1, \cdots, r_M\}$ can be put on a junction tree then $S_R$ has no loops.

The following theorem states necessary and sufficient condition for the Kikuchi approximate free energy and the consistency constraint set of pseudo-marginals to be exact:

**Theorem 7.** (Exactness of Kikuchi approximates, $\Delta^K_R$ and $F^K_R$)

A) $\Delta^K_R = \Delta_R$ iff $S_R$ is loop-free.

B) Let $b(x)$ be a distribution with marginals $b_r(x_r)$. Then

$$F^K_R(\{b_r, r \in R\}) = F(b) \text{ iff } \forall x, \ b(x) = \prod_{r \in R} b_r(x_r)^{c_r}$$

**Corollary 8.** If $S_R$ has no loops, then the constrained minimization problem (11) has a unique solution. Further, the solution $\{b^*_r, r \in R\}$ is the exact marginals of the product function, i.e. $b^*_r(x_r) = (\sum_{x \setminus x_r} \prod_{r \in R} \alpha_r(x_r)) / Z$.

In fact in the case when $S_R$ has no loops, iterative algorithms such as GBP of [9] converge in finite time to the unique solutions $B_r$.

Given a product function of potentials $\{\alpha_r(x_r), r \in R_0\}$ where the regions in $R_0$ cannot be put on a junction tree, it is expected that expanding the collection $R_0$ by adding subsets of $r \in R_0$ as further regions may improve the approximation obtained by the iterative algorithms discussed in Section 4.

However we currently believe that if $S_R$ has loops, for the generic choices of $\alpha_r(x_r)$'s one cannot get \textquote{exact} solutions for all marginals using the Kikuchi approximation method discussed in this paper. This would imply that, given maximal regions $\{r_1, \cdots, r_M\}$, if the usual Belief Propagation algorithm is not expected to give the exact marginals due to existence of loops, then the generalized Kikuchi method for approximation as discussed in this paper is not expected to give the exact marginals either, no matter how many smaller regions are added. Note however that the Kikuchi approximations may be better than those obtained using loopy Belief Propagation. A weaker version of this can be proved easily using Part A) of Theorem 7:
Corollary 9. If $S_R$ has loops, then there exist a collection of potentials $\{\beta_r(x_r), r \in R\}$ such that the constrained minimization problem of (11) has minimizers $\{b_r, r \in R\}$ which are different from the marginals of the product distribution $B(x) = \frac{1}{Z} \prod_{r \in R} \beta_r(x_r)^{c_r}$.

Finally, it can also be shown that for a certain class of collection of regions $R$, (which includes the cluster variational method discussed in [9],) condition (15) of Theorem 1 is equivalent to existence of zero or one loop in $S_R$, i.e. $F^K_R(\{b_r\})$ is convex if $S_R$ has zero or one loop.

In short, as one expects, there is a direct correspondence between fundamental properties of Kikuchi approximation (11) and the graph-theoretic properties of minimal graphical representation $S_R$. Further, iterative algorithms of Section 4, when defined on the minimal graph $S_R$, will involve fewest number of messages compared to any other such algorithm. Although the fixed points of all such algorithms will be identical, it will be interesting to compare the convergence properties of this minimal algorithm to those of GBP algorithms of [10].

References


