Error Bounds Between Marginal Probabilities and Beliefs of Loopy Belief Propagation Algorithm

Nobuyuki Taga and Shigeru Mase

Tokyo Institute of Technology, Ookayama 2-12-1-w8-28,
Meguro-ku, Tokyo 152-8552, Japan
{Nobuyuki.Taga, mase}@is.titech.ac.jp

Abstract. Belief propagation (BP) algorithm has been becoming increasingly a popular method for probabilistic inference on general graphical models. When networks have loops, it may not converge and, even if converges, beliefs, i.e., the result of the algorithm, may not be equal to exact marginal probabilities. When networks have loops, the algorithm is called Loopy BP (LBP). Tatikonda and Jordan applied Gibbs measures theory to LBP algorithm and derived a sufficient convergence condition. In this paper, we utilize Gibbs measure theory to investigate the discrepancy between a marginal probability and the corresponding belief. Consequently, in particular, we obtain an error bound if the algorithm converges under a certain condition. It is a general result for the accuracy of the algorithm. We also perform numerical experiments to see the effectiveness of the result.

1 Introduction

Belief propagation (BP) algorithm has become a popular method of solving inference problems exactly for probabilistic networks without loops (e.g., Bayesian networks) in a finite number of times. It has the origin in the probabilistic expert system theory proposed by Pearl *et al.* [6]. Similar algorithms appear in several applications, such as Viterbi algorithm in hidden Markov models, iterative algorithms for Gallager codes and turbocodes, Kalman filter and the transfer-matrix approach in physics.

It is also widely applied to networks with loops. In that case, the algorithm is called loopy BP (LBP). LBP algorithm, however, may not converge and, even if it does, the solution may not be equal to the target marginal probabilities. Nevertheless, applications of the LBP algorithm are reported to be remarkably good such as in the coding theory (cf. Frey [1], McEliece *et al.* [4] and Murphy *et al.* [5]).

Weiss [9] discussed the LBP algorithm on networks with a single loop and Weiss and Freeman [10] discussed the LBP algorithm on Gaussian networks. A basic idea of Weiss is the fact that the calculation of the LBP algorithm is equivalent to that on a corresponding infinite tree called the computation tree. Tatikonda and Jordan [8] pursued his idea and formulated the convergence

problem as that of Gibbs measures on the computation trees. They showed a relationship between the convergence of LBP algorithm and the phase transition phenomena on the associated computation trees in their paper.

So far, some studies were reported for the general convergence property of LBP algorithm. However, there are few general discussions of its accuracy.

In this paper, we use Gibbs measure theory to measure the discrepancy of marginal probabilities and the corresponding beliefs of LBP algorithm using the concept of the computation tree.

We give a review of the BP algorithm in Sect. 2. In Sect. 3, we introduce Gibbs measure theory and review the application results for LBP algorithm. In Sect. 4, we introduce the concept of measuring discrepancy of two probability measures developed in Gibbs measure theory, apply it to the LBP algorithm with pair potentials, and show some results. In Sect. 5, we report numerical experiments done to see the effectiveness of obtained results. In Sect. 6, we give a conclusion and some remarks.

2 BP Algorithm and Computation Trees

The BP algorithm used in this paper is as follows. Let G be a connected and undirected finite network. Let consider an associated set of random variables $X = \{X_i, i \in G\}$ and its observations $Y = \{y_i, i \in G\}$. The state space E_i of X_i is finite. Some y_i may be missing. We consider a probability function on G of the form

$$p(x \mid y) \equiv P(X = x \mid Y = y) = \frac{1}{Z} \prod_{i \sim j} \phi_{ij}(x_i, x_j) \prod_{i \in G} \phi_i(x_i, y_i) ,$$

where $\sim$ denotes the neighborhood relationship, and the first product extends over all neighboring nodes (i, j). Here $i \in G$ is said to be a neighbor of $j \in G$ if there exists an edge between i and j in G. We call (G, p) a *probabilistic network* with the network G and the joint distribution p. Throughout this paper, Z stands for normalizing constants and are not always the same. Usually, the existence of a data y_i restricts the state space E_i to $\{y_i\}$ effectively. We will adopt this convention and, further, suppress the dependencies of ϕ_i's on $\{y_i\}$. Therefore, it takes the form

$$p(x) = \frac{1}{Z} \prod_{i \sim j} \phi_{ij}(x_i, x_j) \prod_{i \in G} \phi_i(x_i) . \quad (1)$$

It is the basic assumption of this paper that $\phi_{ij}(\cdot, \cdot)$ and $\phi_i(\cdot)$ are all positive.

For each pair of neighboring nodes (i, j) and each state $x_j \in E_j$, we consider the *message* $m_{ij}^{(n)}(x_j)$, $n = 1, 2, \ldots$. These messages obey the following update rule called the *belief propagation* (BP):

$$m_{ij}^{(n+1)}(x_j) = \frac{1}{Z} \sum_{x_i \in E_i} \phi_{ij}(x_i, x_j) \phi_i(x_i) \prod_{k \in \partial i \setminus \{j\}} m_{ki}^{(n)}(x_i) ,$$

where ∂i denotes the set of all neighboring nodes of i. In the following, $|A|$ for set A means its cardinality. All messages are initialized as $m_{ij}^{(0)}(x_j) \equiv 1$. If message $m_{ij}^{(n)}(x_j)$ converges, its limit is denoted by $m_{ij}(x_j)$. For these limit messages, a *belief* for each node i is the normalized product

$$b_i(x_i) = \frac{1}{Z}\phi_i(x_i) \prod_{k \in \partial i} m_{ki}(x_i), \quad x_i \in E_i \ .$$

If a probabilistic network has no loops, i.e., tree-like, it is known that all the messages $\{m_{ij}^{(n)}(x_j)\}$ converge after a finite number of the BP updates and that the belief $b_i(\cdot)$ is equal to the marginal probability $\boldsymbol{P}\{x_i = \cdot\}$ for each $i \in G$, see Jensen [3]. On the other hand, for networks with loops, the messages may not converge and, if converge, the beliefs may not be equal to the marginal probabilities. In particular, for a probabilistic network with loops, this algorithm is called *loopy belief propagation* (LBP). To study the properties of the LBP algorithm, Weiss [9] introduced a concept of *unwrapped networks* (*computation trees* in Tatikonda and Jordan [8]), which are associating infinite trees T_k, $k \in G$. T_k is the limit of increasing finite trees $\{T_k^{(n)}\}$, $n = 1, 2, \ldots$, defined as follows, see Fig. 1.

1. Let $N_i = 0$, $i \neq k$, and $N_k = 1$. For convenience, let $T_k^{(0)} = \{k^{(1)}\}$ where $k^{(1)}$ is a copy of k.
2. Let $\{i, j, \ldots\} = \partial k$, $N_i = N_j = \cdots = 1$ and $i^{(1)}, j^{(1)}, \ldots$ be copies of $i, j, \ldots$ respectively. The first computation tree $T_k^{(1)}$ consists of nodes $k^{(1)}, i^{(1)}, j^{(1)}, \ldots$ and corresponding edges $(k^{(1)}, i^{(1)}), (k^{(1)}, j^{(1)}), \ldots$.
3. If the n-th computation tree $T_k^{(n)}$ is defined, the next computation tree $T_k^{(n+1)}$ is defined to be $T_k^{(n)}$ augmented by new nodes and edges repeating the following steps:
   (a) For each edge $(r^{(\ell)}, s^{(m)})$ of $T_k^{(n)}$ with $r^{(\ell)} \notin T_k^{(n-1)}$, let $i, j, \ldots$, be the nodes $\partial r \backslash \{s\}$ (if non-empty).
   (b) Let $N_i \leftarrow N_i + 1, N_j \leftarrow N_j + 1, \ldots$ and $i^{(N_i)}, j^{(N_j)}, \ldots$, be new copies of $i, j, \ldots$ respectively. Add new nodes $i^{(N_i)}, j^{(N_j)}, \ldots$ and corresponding edges $(i^{(N_i)}, r^{(\ell)}), (j^{(N_j)}, r^{(\ell)}), \ldots$ to $T_k^{(n)}$.

The state space E_i is associated with each node $i^{(n)} \in T_k$ and let $\phi_{i^{(n)}j^{(m)}} = \phi_{ij}$ and $\phi_{i^{(n)}} = \phi_i$. If G has no loops, T_k is the same as G except labeling of nodes. It is easily seen that the message $m_{jk}^{(n)}(x_k)$ which is the result of the n-th BP update with the parallel update rule on G starting from k is equal to $m_{j^{(1)}k^{(1)}}^{T_k^{(n)}}(x_k)$, the result of the n-th BP update of messages performed on $T_k^{(n)}$, that is, on T_k starting from $k^{(1)}$. Therefore, the limiting message heading for k, if exists, is the same for both G and T_k, a key idea why we consider the computation trees besides the original probabilistic networks.

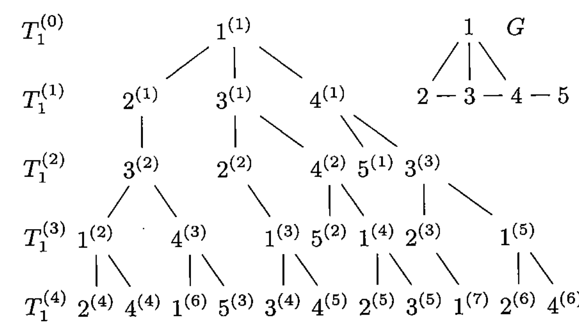

Fig. 1. A network G and the corresponding computation tree for the root node 1 with depth 4

3 Gibbs Measures and LBP Algorithm for Pair Potentials

In this section, we introduce Gibbs measure theory briefly and review the relationships with LBP algorithm.

Let S be a finite or infinite site set. A discrete and finite state space E_i is associated with each $i \in S$. A *configuration* Ω is defined by the set of all possible configurations. Specifically, $\Omega \equiv E^S = \prod_{i \in S} E_i$. Its restriction to a subset $\Lambda \subset S$ is denoted by Ω_Λ. Let $\mathcal{I}$ be the set of non-empty finite subsets of S. A σ-field of Ω is denoted by $\mathcal{F}$. An *interaction potential* (or simply a *potential*) is a family $\Phi = (\Phi_A)_{A \in \mathcal{I}}$ of functions $\Phi_A : \Omega \mapsto \mathbb{R}$ with the following properties; (i) for each $A \in \mathcal{I}$, Φ_A is $\mathcal{F}_A$-measurable. Here $\mathcal{F}_A$ is the restriction of $\mathcal{F}$ to A. (ii) For all $\Lambda \in \mathcal{I}$ and $\omega \in \Omega$, the series $\sum_{A \in \mathcal{I}, A \cap \Lambda \neq \emptyset} \Phi_A(\omega)$ exists.

A *Gibbs specification* for a potential Φ is a system $\{\gamma_\Lambda(\cdot \mid \xi) : \Lambda \in \mathcal{I}, \xi \in E^S\}$ of probability measures defined by

$$\gamma_\Lambda(x|\xi) = \frac{1}{Z_{\Lambda,\xi}} \exp\left\{ -\sum_{A \subset \Lambda} \Phi_A(x_A) - \sum_{A \cap \Lambda \neq \emptyset} \Phi_A(x_{A_\Lambda}, \xi_{A_{\Lambda^c}}) \right\}$$

for all $\Lambda \in \mathcal{I}$ and $x \in E^\Lambda$, where $Z_{\Lambda,\xi}$ is the normalizing constant called the *partition function* and $A_\Lambda = A \backslash \Lambda$. The measure $\gamma_\Lambda(x \mid \xi)$ is called the *Gibbs distribution in Λ with boundary condition ξ*. It is noted that $\gamma_\Lambda(x \mid \xi)$ is dependent on ξ only through $\xi_{\partial\Lambda}$. A probability measure μ on $(E^S, \mathcal{B})$ is called a *Gibbs measure* for Φ if it satisfies the following *DLR* (*Dobrushin-Lanford-Ruelle*) equations:

$$\mu(x \mid \mathcal{B}_{S\backslash\Lambda} = \xi) = \gamma_\Lambda(x \mid \xi), \quad \xi \in E^{\partial\Lambda}, \tag{2}$$

for all $\Lambda \in \mathcal{I}$, where $x \in E^\Lambda$ is canonically embedded into E^S as $x \times E^{S\backslash\Lambda}$. Since $\mu(x \mid \mathcal{B}_{S\backslash\Lambda}) = \mu(x \mid \mathcal{B}_{\partial\Lambda})$, such μ is also called a *Markov random field*.

It should be noted that, for a certain potential Φ, there is a possibility that the Gibbs measure μ which satisfies (2) is not unique. Let $\mathcal{G}_\Phi$ denote the set of all Gibbs measures for a potential Φ. Also, the notation $\mathcal{G}(\gamma)$ for a specification γ is often used in particular when one is conscious of the conditional probabilities rather than the potential. In terms of Gibbs measure theory, it is said that a *phase transition* occurs if $|\mathcal{G}_\Phi| > 1$ (i.e., $|\mathcal{G}(\gamma)| > 1$).

Tatikonda and Jordan [8] applied the theory of Gibbs measures to study the property of the LBP algorithm through the concept of computation tree in pair potential case, i.e., the potential Φ is defined by $\{\Phi_i, \Phi_{ij}\}$ where $\{\Phi_i\}$ and $\{\Phi_{ij}\}$ are certain 1-body and 2-body potentials.

In fact, the properties of Gibbs measures defined on general tree networks had already been discussed in Gibbs measure theory. In that discussion, the concept of *boundary law* is utilized as an important concept. Tatikonda and Jordan showed the relationship between the convergent messages and the boundary law for the associated Gibbs measure on the corresponding limit computation tree. As a result, they concluded that the uniqueness of boundary law guarantees the convergence of the LBP algorithm. They also introduced an uniqueness condition called Simon's condition of Gibbs measure theory as a convergence condition of the LBP algorithm.

Recently, Taga and Mase [7] discussed the difference of convergence ratio between so-called sequential and parallel update orders using Gibbs measure theory. In their paper, they showed sequential update order always converges faster than parallel one under the condition of absence of phase transitions. They also showed sequential update order is expected to converge faster generally through numerical experiments.

4 Comparison Between Marginal Probabilities and Beliefs

We show another application of Gibbs measure theory in this section to measure the discrepancies between marginal probabilities of probabilistic networks with pair potentials and the corresponding beliefs. First, we need to introduce some concepts which can be used for Gibbs measures on general networks. In the following, we only give a brief introductions and reviews of notations. More precisely, see Georgii [2].

Let E and $\mathcal{E}$ be some state space and the arbitrary σ-field respectively. Then $(E, \mathcal{E})$ is a measurable space. Let p_1 and p_2 be two probability measures on $(E, \mathcal{E})$. We define a distance $||p_1 - p_2||$ of p_1 and p_2 by

$$||p_1(\cdot) - p_2(\cdot)|| \equiv \max_{A \in \mathcal{E}} |p_1(A) - p_2(A)| \ .$$

It is clear that $||\cdot||$ is one half of total variation distance. Let S be an arbitrary (not necessarily tree) site set and Ω be a set of all possible configurations on S. Let γ be a specification on Ω. For each pair of sites $i, j \in S$, we define

$$C_{ij}(\gamma) = \sup_{\zeta, \eta \in \Omega, \zeta_{S\setminus\{j\}} = \eta_{S\setminus\{j\}}} ||\gamma_i(\cdot|\zeta) - \gamma_i(\cdot|\eta)|| \ .$$

The matrix $C(\gamma) = (C_{ij}(\gamma))_{i,j \in S}$ is called *Dobrushin's interdependence matrix* for γ. A real function f on Ω is called a *cylinder function* or a *local function* if f is $\mathcal{F}_\Lambda$-measurable for some finite Λ where $\mathcal{F}_\Lambda$ denotes the σ-field of Ω_Λ, i.e., the restriction of Ω to Λ. A function $f : \Omega \mapsto \mathbb{R}$ will be said to be *quasilocal* if there is a sequence $(f_n)_{n \geq 1}$ of local functions f_n such that $\lim_{n\to\infty} \sup_{\omega \in \Omega} |f(\omega) - f_n(\omega)| = 0$. We write $\overline{\mathcal{L}}$ for the set of all bounded quasilocal functions. Let $p(f)$ denote the expectation of f with respect to a probability p. A specification γ is said to be quasilocal if $\gamma_\Lambda(f|\cdot)$ is quasilocal for each $\Lambda \in \mathcal{J}$ and $f \in \overline{\mathcal{L}}$.

We introduce here a well-known condition for absence of phase transition. It is said that a specification γ satisfies *Dobrushin's condition* if γ is quasilocal and

$$c(\gamma) \equiv \sup_{i \in S} \sum_{j \in S} C_{ij}(\gamma) < 1 \ .$$

Let $f \in \overline{\mathcal{L}}$ and $j \in S$ be given. The oscillation of f at j is defined by

$$\delta_j(f) = \sup_{\zeta, \eta \in \Omega, \zeta_{S\setminus\{j\}} = \eta_{S\setminus\{j\}}} |f(\zeta) - f(\eta)| \ . \tag{3}$$

Let $\mathcal{F}$ be a σ-field of Ω. Then we are ready to introduce a tool used to measure discrepancy of two probability measures defined on $(\Omega, \mathcal{F})$. Let two probability measures μ and $\tilde{\mu}$ on $(\Omega, \mathcal{F})$ be given. A vector $a = (a_i)_{i \in S} \in [0, \infty)^S$ is called an *estimate for μ and $\tilde{\mu}$* if

$$|\mu(f) - \tilde{\mu}(f)| \leq \sum_{j \in S} a_j \delta_j(f) \tag{4}$$

for all $f \in \overline{\mathcal{L}}$. We state two basic facts known about the estimates. First, the constant vector $a \equiv (1)_{i \in S}$ is always an estimate. Second, let fix two specifications γ and $\tilde{\gamma}$, and let $\mu \in \mathcal{G}(\gamma)$ and $\tilde{\mu} \in \mathcal{G}(\tilde{\gamma})$ be given. Suppose a is an estimate for μ and $\tilde{\mu}$. Define $\overline{a}_i$ by

$$\overline{a}_i = \sum_{j \in S} C_{ij}(\gamma) a_j + \tilde{\mu}(\beta_i) \tag{5}$$

for every $i \in S$, where $\beta_i : \Omega \to [0, \infty)$ is a measurable function such that

$$||\gamma_i(\cdot|\omega) - \tilde{\gamma}_i(\cdot|\omega)|| \leq \beta_i(\omega) \ . \tag{6}$$

Then $\overline{a} = (\overline{a}_i)_{i \in S}$ is an estimate for μ and $\tilde{\mu}$.

In the following, we try to derive some properties specific to the LBP algorithm. It is noted that the beliefs are, if message update converges, the marginal probabilities of a single site of an associated Gibbs measure on the corresponding computation tree [8]. On the basis of this fact, we look at a certain indicator function f as follows:

Proposition 1. *Fix $x_i \in E_i$ for some $i \in S$. Let $f : \Omega \mapsto \{0, 1\}$ be defined by $f(\omega) = 1_{\{x_i\}}(\omega_i)$. Then $f \in \overline{\mathcal{L}}$, and following two corollaries hold.*

Corollary 1. *$\mu(1_{\{x_i\}})$ is the marginal probability for $X_i = x_i$,*

Corollary 2. *$\delta_j(1_{\{x_i\}}) = 1$ if $j = i$, otherwise 0.*

Proof. First corollary is trivial so that we only show the other. We write a configuration $\omega = \omega_j\omega_{S\backslash\{j\}}$ separating with respect to a site $j \in S$ and the other sites $S\backslash\{j\}$. Fix a site $i \in S$ and $x_i \in E_i$. Then we can write eq. (3) with $f(\omega) = 1_{\{x_i\}}(\omega_i)$ as

$$\begin{aligned}\delta_j(f) &= \sup_{\zeta,\eta\in\Omega\ ,\zeta_{S\backslash\{j\}}=\eta_{S\backslash\{j\}}} |f(\zeta) - f(\eta)| \\ &= \sup_{\omega\in\Omega}\sup_{x,y\in E_j} |f(x\omega_{S\backslash\{j\}}) - f(y\omega_{S\backslash\{j\}})| \\ &= \begin{cases} \sup_{x,y\in E_i} |1_{\{x_i\}}(x) - 1_{\{x_i\}}(y)| & \text{if } j = i\ , \\ \sup_{\omega_j\in E_j} |1_{\{x_i\}}(\omega_j) - 1_{\{x_i\}}(\omega_j)| & \text{otherwise}\ , \end{cases} \\ &= \begin{cases} 1 \text{ if } j = i\ , \\ 0 \text{ otherwise}\ . \end{cases}\end{aligned}$$

The proof is thus complete. □

We think of μ and $\tilde{\mu}$ shown above as the probability of a target probability and the associated Gibbs measure and try to measure the discrepancy between their the marginal probabilities. The following two propositions are necessary for this.

Proposition 2. *Let G be the network of a probabilistic network and T be the associated computation tree. Assume G' is the network such that $G' = \{i^{(1)}; i \in G\}$ and G' has an edge between $i^{(1)}$ and $j^{(1)}$ if there exists an edge between i and j in G. Then one can construct a certain network S such that $G' \subset S$ and $T \subset S$. We will call S a common space of G and T.*

Proof. Let B_T denote the edge set of T. There exist edges such that $k^{(1)}l^{(n)}$ for $n \geq 2$ in B_T, i.e., the neighboring sites such that one of it has 1 as superscript and another has $n > 1$ as superscript. For every such $k^{(1)}l^{(n)}$, if $k^{(1)}l^{(1)} \notin B_T$, add $k^{(1)}l^{(1)}$ to B_T. The resulting site set (T, B_T) is the common space S. □

We give an example of a common space in Figure 2.

Proposition 3. *Suppose the joint distribution p of a probabilistic network (G,p) has the form (1). Let $\Phi_i = \log\phi_i$, $\Phi_{ij} = \log\phi_{ij}$ for $i,j \in G$. Let B_G, B_T be the edge set of G and the corresponding computation tree T. We define two interaction potentials Φ and $\tilde{\Phi}$ for the common space S of G and T as follows.*

$$\begin{aligned}\Phi &\equiv \{\Phi_{i^{(1)}} = \Phi_i; i \in G\} \cup \{\Phi_{i^{(1)}j^{(1)}} = \Phi_{ij}; ij \in B_G\} \\ &\quad \cup \{\Phi_{i^{(k)}} = \Phi_{i^{(k)}j^{(l)}} = 0; k \text{ or } l \geq 2\}\ , \\ \tilde{\Phi} &\equiv \{\tilde{\Phi}_{i'} = \Phi_i; i' \in T\} \cup \{\tilde{\Phi}_{i'j'} = \Phi_{ij}; i'j' \in B_T\} \cup \{\tilde{\Phi}_{i'j'} = 0; i'j' \notin B_T\}\ ,\end{aligned}$$

where ij stands for the edge between i and j. Then the systems of conditional probabilities for Φ and $\tilde{\Phi}$ are the Gibbs specifications defined on S.

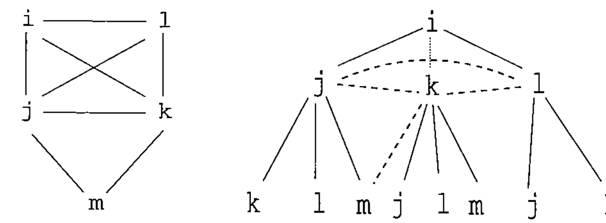

Fig. 2. A network G (left) and the common space S (right) with depth 2 from the root node i. The superscripts of the indices in S are ignored. Dotted lines correspond to the lines added to the computation tree T in the construction of S.

It should be noted that the transformation of potentials makes no difference to the marginal probabilities for original space; in particular, for $\mu \in \mathcal{G}(\Phi)$ and $\tilde{\mu} \in \mathcal{G}(\tilde{\Phi})$, $\mu(x_i)$ and $\tilde{\mu}(x_i)$ for each $i \in G'$ are equal to $p(x_i)$ of the target probability and the corresponding belief (if exists) respectively.

We let γ and $\tilde{\gamma}$ denote the specifications for above Φ and $\tilde{\Phi}$ respectively. γ and $\tilde{\gamma}$ have the following property.

Corollary 3. *There exists a non-empty set S' of indices $i \in S$ such that*

$$\gamma_i(x_i|\omega) = \tilde{\gamma}_i(x_i|\omega) \tag{7}$$

for all $x_i \in E_i$ and $\omega \in \Omega$.

Proof. In deed, the node in S corresponds to the root node of the computation tree is such a site. The other $i \in S$ can be such a site if i is originated from G and all the edges connecting with i in S are originated from both T and G. Each such site can be shown to satisfy (7) by direct calculations of conditional probabilities of two specifications. □

According to the above property, it is clear that

$$||\gamma_i(\cdot|\omega) - \tilde{\gamma}_i(\cdot|\omega)|| = 0, i \in S'\ ,$$

for all $\omega \in \Omega$. Thus we can put $\beta_i(\cdot) \equiv 0$ in (6) for each $i \in S'$. We are now ready to give the following results.

Theorem 1. *Let $b_i(\cdot)$ be a convergent belief for $i \in G$ of a probabilistic network (G,p). Let γ be the specification corresponding to the probabilistic network and $C(\gamma)$ be the Dobrushin's interdependence matrix for γ. Define $c_i(\gamma) = \sum_{j\in G} C_{ij}(\gamma)$. Then*

$$|p(x_i) - b_i(x_i)| \leq \min\{1, c_i(\gamma)\}$$

for all $x_i \in E_i$.

Proof. We consider the computation tree T with the root node i. Suppose S is the common space of G and T. Let $\tilde{\gamma}$ be the specification corresponding to the associated computation tree. Let $\mathcal{G}(\tilde{\gamma})$ be the set of all Gibbs measures for $\tilde{\gamma}$ and fix a Gibbs measure $\tilde{\mu} \in \mathcal{G}(\tilde{\gamma})$ for $\tilde{\gamma}$. With $f(\omega) = 1_{\{x_i\}}(\omega_i)$ and Corollary 2, we can write eq. (4) as

$$|\mu(1_{\{x_i\}}) - \tilde{\mu}(1_{\{x_i\}})| \leq a_i \ .$$

for some Gibbs measure $\mu \in \mathcal{G}(\gamma)$ for γ. Using the trivial estimate $a = (1)_{i \in S}$, we can obtain $\overline{a}_i$ from eq. (5) as

$$\overline{a}_i = \sum_{j \in S} C_{ij}(\gamma) + \tilde{\mu}(\beta_i) = \sum_{j \in G'} C_{ij}(\gamma) = c_i(\gamma) \ .$$

Here we took $\beta_i(\omega) \equiv 0$ since i is the root node of computation tree. The second equation comes from the fact that $C_{ij}(\gamma) = 0$ for i, j, such that $i \not\sim j$ for γ, and i is the site at which (7) is satisfied. It should be noted that a' such that $a'_i \equiv \min\{a_i, \overline{a}_i\} = \min\{1, c_i(\gamma)\}$ and $a'_k \equiv 1$ for $k \neq i$ is also an estimate. The marginal probabilities μ and $\tilde{\mu}$ are in fact that of p and the belief for the node corresponding to root node respectively. Thus the proof is complete. □

If there is at least one site i such that $c_i(\gamma) < 1$, its factor of an estimate a_i can be taken less than 1. On the other hand, when a site j has a neighbor i such that $a_i < 1$, its factor of the estimate a_j may be taken less than 1 even if $c_j(\gamma) \geq 1$ using (5) with $a_i < 1$. Conversely, if a_j decreases, a_i becomes smaller using (5) with a_j again. Such a mutual improvement can be utilized below.

Corollary 4. *Let γ and $\tilde{\gamma}$ be the specifications corresponding to a probabilistic network and the corresponding computation tree defined on a certain common space respectively. Let $C(\gamma)$ be the Dobrushin's independence matrix for the specification γ and $\tilde{\mu} \in \mathcal{G}(\tilde{\gamma})$. Let $a^{(n)} = (a_i^{(n)})_{i \in S}, n = 1, 2, \ldots,$. be defined by*

$$a_i^{(n+1)} = \min\{1, a_i^{(n)}, (C(\gamma)a^{(n)} + \tilde{\mu}(\beta))_i\}, \ i \in S \ , \tag{8}$$

where $a_i^{(0)} = 1, i \in S$, and $\tilde{\mu}(\beta) = (\tilde{\mu}(\beta_i))_{i \in S}$. Then $a^{(n)}$ has a limit a^* and each error bound between the marginal probability and the belief for i is given by a_i^*.

Proof. For each i, $a_i^{(n)}$ clearly does not increase with n. It is also clear that $a_i^{(n)}$ has a lower bound 0 since $a_i^{(0)} = 1$, $i \in S$, and all factors of $C(\gamma)$ and $\tilde{\mu}(\beta)$ are non-negative. Then $a^{(n)}$ has a limit. The result follows from the fact that each $a^{(n)}$ can be an estimate. □

5 Numerical Experiments

In the preceding section, we showed error bounds between marginal probabilities and the corresponding beliefs. In this section, we give numerical experiments so as to see the effectiveness of error bounds based on Theorem 1 and Corollary 4.

We now use following *Ising models* on complete graph with four vertices.

$$p(x) \propto \exp\left(h \sum_{i=1,2,3,4} x_i + J \sum x_i x_j\right)$$

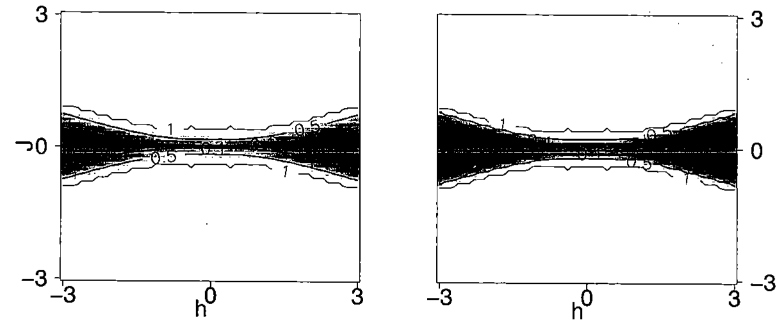

Fig. 3. Error bounds between one-variable marginal probabilities and the corresponding beliefs for Ising models on the complete graph network with four vertices. The left (right) figure is obtained using Theorem 1 (Corollary 4). Contours are imposed.

Here the second summation is taken with respect to all pairs of $\{1, 2, 3, 4\}$. The corresponding computation tree is called the *Cayley tree* of degree 2 in Gibbs measure theory and the associated Gibbs measures are Ising models on it. Let $CT(2)$ denote the Cayley tree of degree 2. For Ising models on Cayley trees, all factors of the Dobrushin's interdependence matrix are same and it is easy to calculate. In particular, for $CT(2)$, the constant $c(h, J)$ for each h, J is written by

$$C_{ij}(\gamma) = \frac{\sinh(2|J|)}{g(h, J) + \cosh(2J)} \equiv c(h, J) \ ,$$

where $g(x, y) = \cosh 2(|x| + |y|)$ if $|x| \leq |y|$, otherwise $\cosh 2(|x| - 2|y|)$. Then $c_i(\gamma)$ shown in Theorem 1 will be $3c(h, J)$. In calculation of (8), we need to fix $\beta(\omega) = (\beta_i(\omega))_{i \in CT(2)}$ and to obtain the expectation $\tilde{\mu}(\beta) = (\tilde{\mu}(\beta_i))_{i \in CT(2)}$. The left side of (6) is clearly bounded by 1, so that we put here $\beta_i(\omega) \equiv 1$ for each i. Then all the factors of the expectation $\tilde{\mu}(\beta)$ will be 1; we use this in calculation of (8). In Fig. 3, we summarize the results.

Let $B_{h,J}$, $B^*_{h,J}$ be the error bounds obtained based on Theorem 1 and Corollary 4 for each (h, J) respectively. For all h and J, it is shown that $B^*_{h,J} \leq B_{h,J}$. This means that the use of (8) are effective for obtaining better results in this case. Nevertheless, seen from the experimental results for Ising models reported in [7], the region where one can get the good error bound is restrictive. It should be noted that the region (J, h) where $B_{h,J} < 1$ is very close to the region where Dobrushin's condition is satisfied.

6 Conclusion and Remarks

We applied Gibbs measure theory to LBP algorithm with pair potentials to obtain error bounds between marginal probabilities and the corresponding beliefs. We showed a nontrivial error bound can be obtained under a certain condition for each site if the algorithm converged. We also gave a procedure which has a potential for improving the error bounds. We gave numerical experiments to check the effectiveness. In some cases, such as Dobrushin's condition is satisfied, the error bounds and the improvement procedure seem effective. Nevertheless, the region where one can obtain good bounds seems restrictive.

We give some remarks in the rest of this section. First, the concept of estimates we used in this paper was developed for general Gibbs measures, so that there may be a possibility of improvements in application to LBP. Second, we used 1 as the factors of $\tilde{\mu}(\beta)$ in the numerical experiments. However, the precise assessment of $\tilde{\mu}(\beta)$ surely has an influence for obtaining a good error bound. The last remark is about higher-order potential case. In fact, there is another version of LBP algorithm called *LBP on factor graphs*, with which one can treat higher-order potentials. Similar to the pair potential case introduced in this paper, one can think the concept of computation trees for LBP on factor graphs. Under the computation tree for LBP on factor graphs, the result shown in this paper would be valid for probability function with higher-order potentials.

References

1. Frey, B. J.: Graphical Models for Pattern Classification, Data Compression and Channel Coding, MIT press, Cambridge (1998).
2. Georgii, H. -O.: Gibbs Measures and Phase Transitions, Walter de Gruyter, Berlin · New York (1988).
3. Jensen, F.: An Introduction to Bayesian Networks, UCL Press, London (1996).
4. McEliece, R. J., MacKay, D. J. C., Cheng, J. F.: Turbo Decoding as an Instance of Pearl's "Belief Propagation" Algorithm, IEEE Journal on Selected Areas in Communication, **16(2)** Springer Verlag, New York, Berlin, Heidelberg (1998) 140-152.
5. Murphy, K. P., Weiss, Y., Jordan, M. I.: Loopy belief propagation for approximate inference: an empirical study, Proc. of the 15th Conf. on Unc. in Art. Int., Morgan Kaufmann, San Francisco (1999) 467-475.
6. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann, San Francisco (1988).
7. Taga, N., Mase, S.: On the Convergence of Loopy Belief Propagation Algorithm for Different Update Rules. IEICE transactions on Fundamentals of Electronics, Communications and Computer Sciences, Vol.E89-(2), Tokyo (2006) 575-582.
8. Tatikonda, S. C., Jordan, M. I.: Loopy Belief Propagation and Gibbs Measures, Proc. of the 18th Conf. on Unc. in Art. Int., Morgan Kaufmann, San Francisco (2002) 493-500.
9. Weiss, Y.: Correctness of Local Probability Propagation in Graphical Models with Loops, Neur. Comp., vol.12, (2000) 1-41.
10. Weiss, Y., Freeman, W. T.: Correctness of Belief Propagation in Gaussian Graphical Models of Arbitrary Topology, Neur. Comp., vol.12, (2001) 2173-2200.

Applications of Gibbs Measure Theory to Loopy Belief Propagation Algorithm

Nobuyuki Taga and Shigeru Mase

Tokyo Institute of Technology, Ookayama 2-12-1-w8-28,
Meguro-ku, Tokyo 152-8552, Japan
{Nobuyuki.Taga, mase}@is.titech.ac.jp

Abstract. In this paper, we pursue application of Gibbs measure theory to LBP in two ways. First, we show this theory can be applied directly to LBP for factor graphs, where one can use higher-order potentials. Consequently, we show beliefs are just marginal probabilities for a certain Gibbs measure on a computation tree. We also give a convergence criterion using this tree. Second, to see the usefulness of this approach, we apply a well-known general condition and a special one, which are developed in Gibbs measure theory, to LBP. We compare these two criteria and another criterion derived by the best present result. Consequently, we show that the special condition is better than the others and also show the general condition is better than the best present result when the influence of one-body potentials is sufficiently large. These results surely encourage the use of Gibbs measure theory in this area.

1 Introduction

Inference problems using graphical models are important in various application fields. The belief propagation (BP) algorithm is an efficient method for computing marginal probabilities of probabilistic networks without loops. BP can be formally applied also to networks with loops (LBP). However, if networks have loops, the algorithm may not converge and beliefs may not equal to exact marginal probabilities. Nevertheless, applications of LBP algorithm have been reported to be remarkably useful such as in the coding theory [1,4,6].

In analysis of LBP, Tatikonda and Jordan [8] applied Gibbs measure theory using the concept of computation trees, which was first introduced by Weiss [9]. They also gave a sufficient convergence criterion based on Simon's condition of Gibbs measure theory. Nevertheless, to use this theory seems not to be so popular.

In this paper, we pursue Gibbs measure approach. This paper is composed of two parts. First, we show that this theory can directly be applied to general potentials case. The concept of computation tree is important to apply Gibbs measure theory to LBP. However, it is discussed only for pair potential case and it is still unclear how to construct it where higher-order potentials exist. We give a construction of computation trees according to the LBP for factor graphs. Second, we show the effectiveness of Gibbs measure approach. Tatikonda and Jordan derived a criterion based on Simon's condition of Gibbs measures theory