2nd International Symposium on
Information Geometry and its Applications
December 12-16, 2005, Tokyo
Pages 000–000

INFORMATION GEOMETRY OF PROPAGATION ALGORITHMS AND APPROXIMATE INFERENCE

SHIRO IKEDA

1. INTRODUCTION

Consider the inference problem of undirected graphical models[8, 9]. When the graph is tree, the Belief Propagation (BP) algorithm (J.Pearl[11]) is an efficient algorithm and the exact inference is computed. However, when the graph is "loopy," and the loops are big, the exact inference becomes intractable.

Besides sampling methods, such as MCMC, tractable approximate inference gives us one practical solution, and the "loopy BP" algorithm is one of the most successful methods. Recently, it is pointed out that the idea of loopy BP have been used many fields, for example, Bethe approximation[3] in statistical physics, and the decoding algorithms of Low Density Parity Check (LDPC) codes [4] and turbo codes [2] in error correction codes.

Although we observe the loopy BP works well in many applications, its theoretical aspects are not fully understood. Among some theoretical studies of loopy BP, we have studied it from information geometrical viewpoint[6, 7]. In this abstract, we first summerize the results of [6, 7] by defining the problem and showing the properties loopy BP based on information geometry[1]. We further show its relation to other propagation algorithms, including the convex concave computational procedure (CCCP)[14] and Adaptive TAP approximation[10]

2. Information Geometry of BP

2.1. **Problem, Family of Distributions, and Projection.** Let $\boldsymbol{x} = (x_1, \cdots, x_n)^T$ be random variables. We consider the case where each x_i is binary i.e., $x_i \in \{-1, +1\}$ for simplicity. The joint distribution of $\boldsymbol{x}$ is $q(\boldsymbol{x})$, and we define the expectation of $\boldsymbol{x}$ as $\hat{\boldsymbol{\eta}}$

$$\hat{\boldsymbol{\eta}} = E_q[\boldsymbol{x}] = \sum_{\boldsymbol{x}} \boldsymbol{x} q(\boldsymbol{x}).$$

In this article, we focus on an inference problem of $\hat{\boldsymbol{\eta}}$, which is equivalent to the inference of $\prod_{i=1}^n q(x_i)$. In graphical models, $q(\boldsymbol{x})$ is often defined as the product of functions $\{\phi_i(x_i)\}$ and clique functions $\{\phi_r(\boldsymbol{x}_r)\}$ as,

$$q(\boldsymbol{x}) = \frac{1}{Z_q} \prod_{i=1}^n \phi_i(x_i) \prod_{r \in \mathcal{C}} \phi_r(\boldsymbol{x}_r) = \exp\Big[\boldsymbol{h}\cdot\boldsymbol{x} + \sum_{r=1}^L c_r(\boldsymbol{x}) - \psi_q\Big], \ \phi_i(x_i) > 0, \ \phi_r(\boldsymbol{x}_r) > 0,$$

here $\mathcal{C}$ is the set of cliques, and L is the cardinality of $\mathcal{C}$.

$$h_i = \frac{1}{2} \ln \frac{\phi_i(x_i = +1)}{\phi_i(x_i = -1)}, \quad c_r(\boldsymbol{x}) = \ln \phi_r(\boldsymbol{x}_r), \quad \psi_q = \ln Z_q.$$

Let us consider the following family of probability distributions

$$S = \Big\{ p(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{v}) = \exp[\boldsymbol{\theta} \cdot \boldsymbol{x} + \boldsymbol{v} \cdot \boldsymbol{c}(\boldsymbol{x}) - \psi(\boldsymbol{\theta}, \boldsymbol{v})] \;\Big|\; \boldsymbol{\theta} \in \Re^n, \boldsymbol{v} \in \Re^L \Big\},$$

$$\boldsymbol{c}(\boldsymbol{x}) = (c_1(\boldsymbol{x}), \cdots, c_L(\boldsymbol{x}))^T, \quad \boldsymbol{v} \cdot \boldsymbol{c}(\boldsymbol{x}) = \sum_{r=1}^{L} v^r c_r(\boldsymbol{x}),$$

where its natural parameter is $(\boldsymbol{\theta}, \boldsymbol{v})$, $\boldsymbol{\theta} = (\theta^n, \cdots, \theta^n)^T$, $\boldsymbol{v} = (v^1, \cdots, v^L)^T$, and clearly $q(\boldsymbol{x}) = p(\boldsymbol{x}; \boldsymbol{h}, \boldsymbol{1}) \in S$. We define M_0 as a submanifold of S specified by $\boldsymbol{v} = \boldsymbol{0}$,

$$M_0 = \Big\{ p_0(\boldsymbol{x}; \boldsymbol{\theta}) = \exp[\boldsymbol{h} \cdot \boldsymbol{x} + \boldsymbol{\theta} \cdot \boldsymbol{x} - \psi_0(\boldsymbol{\theta})] \;\Big|\; \boldsymbol{\theta} \in \Re^n \Big\}.$$

The product of marginal distributions of $q(\boldsymbol{x})$ is included in M_0, that is, $\prod_{i=1}^n q(x_i) \in M_0$. Next we show the definition of the m–projection[1] to M_0.

Definition 1. Let π_{M_0} be the operator of the m–projection to M_0 as follows.

$$\pi_{M_0} \circ r(\boldsymbol{x}) = \operatorname*{argmin}_{\boldsymbol{\theta}} D[r(\boldsymbol{x}); p_0(\boldsymbol{x}; \boldsymbol{\theta})].$$

Here, $D[\cdot;\cdot]$ is the KL (Kullback-Leibler)-divergence defined as

$$D[r(\boldsymbol{x}); p(\boldsymbol{x})] = \sum_{\boldsymbol{x}} r(\boldsymbol{x}) \ln \frac{r(\boldsymbol{x})}{p(\boldsymbol{x})}.$$

Since M_0 is e–flat from its definition, the m–projection is unique, and we observe that the inference of $\hat{\boldsymbol{\eta}}$ is equivalent to compute $\hat{\boldsymbol{\theta}} = \pi_{M_0} \circ q(\boldsymbol{x})$, since $\hat{\boldsymbol{\eta}} = E_{p_0(\boldsymbol{x}; \hat{\boldsymbol{\theta}})}[\boldsymbol{x}]$.

2.2. **Information Geometry of BP.** Although the exact inference is simply denoted as $\hat{\boldsymbol{\theta}} = \pi_{M_0} \circ q(\boldsymbol{x})$, the computation becomes intractable for large loopy graphs. In many applications, we face with the same problem, and the BP algorithm is widely used. For loopy graphs, the BP algorithm does not necessarily converge, and even if it does, the result is not equivalent to the exact inference.

The following submanifold M_r plays an important role to understand BP

$$M_r = \Big\{ p_r(\boldsymbol{x}; \boldsymbol{\zeta}_r) = \exp\Big[\boldsymbol{h} \cdot \boldsymbol{x} + c_r(\boldsymbol{x}) + \boldsymbol{\zeta}_r \cdot \boldsymbol{x} - \psi_r(\boldsymbol{\zeta}_r)\Big] \;\Big|\; \boldsymbol{\zeta}_r \in \Re^n \Big\}, \quad r = 1, \cdots, L.$$

M_r is an e–flat submanifold of S, and its natural parameter is $\boldsymbol{\zeta}_r$. We give the information geometrical view of BP. The well-known definition of BP is found somewhere else [9, 11, 12].

Information geometrical view of BP

(1) *Set* $t = 0$, $\boldsymbol{\zeta}_r^t = \boldsymbol{0}$, $r = 1, \cdots, L$.

(2) *Increment* t *by one and compute* $\boldsymbol{\theta}$ *and* $\{\boldsymbol{\zeta}_r\}$ *iteratively as follows*

$$\begin{aligned} \boldsymbol{\theta}^{t+1} &= \sum_r [\pi_{M_0} \circ p_r(\boldsymbol{x}; \boldsymbol{\zeta}_r^t) - \boldsymbol{\zeta}_r^t], \\ \boldsymbol{\zeta}_r^{t+1} &= \boldsymbol{\theta}^{t+1} - [\pi_{M_0} \circ p_r(\boldsymbol{x}; \boldsymbol{\zeta}_r^t) - \boldsymbol{\zeta}_r^t], \quad r = 1, \cdots, L. \end{aligned}$$

(3) *Repeat step* (2) *until convergence.*

Let the converged point of BP be $\{\boldsymbol{\zeta}_r^*\}$ and $\boldsymbol{\theta}^*$, where $\boldsymbol{\theta}^* = \sum_r \boldsymbol{\zeta}_r^*/(L-1)$. We also define $\boldsymbol{\xi}_r^* = \boldsymbol{\theta}^* - \boldsymbol{\zeta}_r^*$. The probability distribution of $q(\boldsymbol{x})$, its final approximations $p_0(\boldsymbol{x}; \boldsymbol{\theta}^*) \in M_0$, and $p_r(\boldsymbol{x}; \boldsymbol{\zeta}_r^*) \in M_r$ are described as follows

$$\begin{aligned} q(\boldsymbol{x}) &= \exp[\boldsymbol{h} \cdot \boldsymbol{x} + c_1(\boldsymbol{x}) + \cdots + c_r(\boldsymbol{x}) + \cdots + c_L(\boldsymbol{x}) - \psi_q], \\ p_0(\boldsymbol{x}; \boldsymbol{\theta}^*) &= \exp[\boldsymbol{h} \cdot \boldsymbol{x} + \boldsymbol{\xi}_1^* \cdot \boldsymbol{x} + \cdots + \boldsymbol{\xi}_r^* \cdot \boldsymbol{x} + \cdots + \boldsymbol{\xi}_L^* \cdot \boldsymbol{x} - \psi_0(\boldsymbol{\theta}^*)], \\ p_r(\boldsymbol{x}; \boldsymbol{\zeta}_r^*) &= \exp[\boldsymbol{h} \cdot \boldsymbol{x} + \boldsymbol{\xi}_1^* \cdot \boldsymbol{x} + \cdots + c_r(\boldsymbol{x}) + \cdots + \boldsymbol{\xi}_L^* \cdot \boldsymbol{x} - \psi_r(\boldsymbol{\zeta}_r^*)]. \end{aligned}$$

The idea of BP is to approximate $c_r(\boldsymbol{x})$ by $\boldsymbol{\xi}_r^* \cdot \boldsymbol{x}$ in M_r, taking the information from $M_{r'}$ $(r' \neq r)$ into account. The information is integrated in $\boldsymbol{\theta}^*$ in M_0.

The following theorem [6] characterizes the equilibrium of BP.

Theorem 1. *The equilibrium* $(\boldsymbol{\theta}^*, \{\boldsymbol{\zeta}_r^*\})$ *satisfies*

1) *m–condition:* $\boldsymbol{\theta}^* = \pi_{M_0} \circ p_r(\boldsymbol{x}; \boldsymbol{\zeta}_r^*)$.

2) *e–condition:* $\boldsymbol{\theta}^* = \dfrac{1}{L-1} \sum_{r=1}^{L} \boldsymbol{\zeta}_r^*$.

We define two submanifolds M^* and E^* of S as follows,

$$M^* = \Big\{ p(\boldsymbol{x}) \;\Big|\; p(\boldsymbol{x}) \in S,\ \sum_{\boldsymbol{x}} \boldsymbol{x} p(\boldsymbol{x}) = \sum_{\boldsymbol{x}} \boldsymbol{x} p_0(\boldsymbol{x}; \boldsymbol{\theta}^*) = \boldsymbol{\eta}_0(\boldsymbol{\theta}^*) \Big\},$$

$$E^* = \Big\{ p(\boldsymbol{x}) = C p_0(\boldsymbol{x}; \boldsymbol{\theta}^*)^{t_0} \prod_{r=1}^{L} p_r(\boldsymbol{x}; \boldsymbol{\zeta}_r^*)^{t_r} \;\Big|\; \sum_{r=0}^{L} t_r = 1,\ t_r \in \Re \Big\},$$

C : normalization factor.

Note that M^* and E^* are an m–flat and an e–flat submanifold, respectively.

The geometrical implications of the 2 conditions are as follows:

- m–**condition:** The m–flat submanifold M^* which includes $p_r(\boldsymbol{x}; \boldsymbol{\zeta}_r^*)$, $r = 1, \cdots, L$, and $p_0(\boldsymbol{x}; \boldsymbol{\theta}^*)$ is orthogonal to M_r, $r = 1, \cdots, L$ and M_0, that is, they are the m–projections to each other.
- e–**condition:** The e–flat submanifold E^* includes $p_0(\boldsymbol{x}; \boldsymbol{\theta}_0^*)$, $p_r(\boldsymbol{x}; \boldsymbol{\zeta}_r^*)$, $r = 1, \cdots, L$, and $q(\boldsymbol{x})$.

For tree graph, we have the following proposition.

Proposition 1. When $q(\boldsymbol{x})$ is represented with a tree graph, $q(\boldsymbol{x})$, $p_0(\boldsymbol{x}; \boldsymbol{\theta}^*)$, and $p_r(\boldsymbol{x}; \boldsymbol{\zeta}_r^*)$, $r = 1, \cdots, L$ are included in M^* and E^* simultaneously.

This shows when a graph is a tree, $q(\boldsymbol{x})$ and $p_0(\boldsymbol{x}; \boldsymbol{\theta}^*)$ are included in M^* and the fixed point of BP is the exact inference. In the case of a loopy graph, generally $q(\boldsymbol{x}) \notin M^*$ and the exact inference is not a fixed point of BP.

2.3. **Approximate Inference.** However, we still hope that BP gives a good approximation. The difference between the exact inference and the BP solution is regarded as the discrepancy between E^* and M^*. We have given a preliminary analysis in [5, 6], which showed the principal term of the error is directly related to the e–curvature of M^*.

Theorem 2. *Let* $\boldsymbol{\eta}(\boldsymbol{\theta}, \boldsymbol{v}) = E_{p(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{v})}[\boldsymbol{x}]$, *then,* $\hat{\boldsymbol{\eta}} = E_{q(\boldsymbol{x})}[\boldsymbol{x}]$ *is approximated by the decoding result* $\boldsymbol{\eta}(\boldsymbol{\theta}^*, \mathbf{o}) = E_{p_0(\boldsymbol{x}; \boldsymbol{\theta}^*)}[\boldsymbol{x}]$ *as follows*

$$\hat{\boldsymbol{\eta}} \simeq \boldsymbol{\eta}(\boldsymbol{\theta}^*, \mathbf{o}) + \frac{1}{2} \sum_{r \neq s} B_r B_s \boldsymbol{\eta}(\boldsymbol{\theta}, \boldsymbol{v})|_{(\boldsymbol{\theta}, \boldsymbol{v}) = (\boldsymbol{\theta}^*, \mathbf{o})}.$$

where

$$B_r = \frac{\partial}{\partial v^r}\Big|_{\boldsymbol{v}=\mathbf{o}} - \sum_i \tilde{g}_{ir}(\boldsymbol{\theta}^*) \frac{\partial}{\partial \theta^i}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}, \quad \tilde{G}_{\boldsymbol{\theta}\boldsymbol{v}}(\boldsymbol{\theta}) = \{\tilde{g}_{ir}(\boldsymbol{\theta}^*)\} = -\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{v}}.$$

2.4. **Free Energy.** The following free energy is important role to understand the loopy BP

$$\mathcal{F}(\boldsymbol{\theta}, \boldsymbol{\zeta}_1, \cdots, \boldsymbol{\zeta}_L) = (L-1)\psi_0(\boldsymbol{\theta}) - \sum_r \psi_r(\boldsymbol{\zeta}_r) + \Big[\sum_r \boldsymbol{\zeta}_r - (L-1)\boldsymbol{\theta}\Big] \cdot \boldsymbol{\eta}_0(\boldsymbol{\theta}). \tag{1}$$

The above function is rewritten in terms of the KL–divergence as,

$$\mathcal{F}(\boldsymbol{\theta}, \boldsymbol{\zeta}_1, \cdots, \boldsymbol{\zeta}_L) = D[p_0(\boldsymbol{x}; \boldsymbol{\theta}); q(\boldsymbol{x})] - \sum_r D[p_0(\boldsymbol{x}; \boldsymbol{\theta}); p_r(\boldsymbol{x}; \boldsymbol{\zeta}_r)] + C,$$

where C is a constant. The following theorem is easily derived.

Theorem 3. *The equilibrium* $(\boldsymbol{\theta}^*, \boldsymbol{\zeta}_r^*)$ *of BP is a critical point of* $\mathcal{F}(\boldsymbol{\theta}, \boldsymbol{\zeta}_1, \cdots, \boldsymbol{\zeta}_r)$.

Proof. By calculating

$$\frac{\partial \mathcal{F}}{\partial \boldsymbol{\zeta}_r} = \mathbf{o},$$

we have the m–condition, that is

$$\sum_{\boldsymbol{x}} \boldsymbol{x} p_r(\boldsymbol{x}; \boldsymbol{\zeta}_r) = \sum_{\boldsymbol{x}} \boldsymbol{x} p_0(\boldsymbol{x}; \boldsymbol{\theta}),$$

By calculating

$$\frac{\partial \mathcal{F}}{\partial \boldsymbol{\theta}} = \mathbf{o},$$

we are led to the e–condition $(L-1)\boldsymbol{\theta} = \sum_r \boldsymbol{\zeta}_r$. □

2.5. **Local Stability of Fixed Points.** Let $I_0(\boldsymbol{\theta})$ be the Fisher information matrix of $p_0(\boldsymbol{x}; \boldsymbol{\theta})$, and $I_r(\boldsymbol{\zeta}_r)$ be that of $p_r(\boldsymbol{x}; \boldsymbol{\zeta}_r)$, $r = 1, \cdots, L$. Since they belong to the exponential family, we have the following relations:

$$I_0(\boldsymbol{\theta}) = \partial_{\boldsymbol{\theta}\boldsymbol{\theta}} \psi_0(\boldsymbol{\theta}), \quad I_r(\boldsymbol{\zeta}_r) = \partial_{\boldsymbol{\zeta}_r \boldsymbol{\zeta}_r} \psi_r(\boldsymbol{\zeta}_r) \; r = 1, \cdots, L.$$

The local stability of the fixed points of BP algorithm, is influenced by the updating order in the step (2). We show the results when we update all the $\boldsymbol{\zeta}_r$, $r = 1, \cdots, L$ simultaneously[6].

Theorem 4. *Let us define* T *as follows*

$$T = \begin{pmatrix} O & I_0(\boldsymbol{\theta})^{-1} I_2(\boldsymbol{\zeta}_2) - E_n & \cdots & I_0(\boldsymbol{\theta})^{-1} I_L(\boldsymbol{\zeta}_L) - E_n \\ I_0(\boldsymbol{\theta})^{-1} I_1(\boldsymbol{\zeta}_1) - E_n & O & & \vdots \\ \vdots & & \ddots & \vdots \\ I_0(\boldsymbol{\theta})^{-1} I_1(\boldsymbol{\zeta}_1) - E_n & \cdots & \cdots & O \end{pmatrix},$$

where E_n *is the* n *dimensional identity matrix. When* $|\lambda_i| < 1$ *for all* i*, where* λ_i *are the eigenvalues of the matrix* T*, the equilibrium point is locally stable.*

3. Relation to Other Propagation Algorithms

3.1. **CCCP.** Since the equilibrium of BP is characterized with the e– and the m–conditions, there are two naive algorithms to find the equilibrium. One is to constrain the parameters always to satisfy the e–condition, and search for the parameters which satisfy the m–condition (e–constraint algorithm), the other is to constrain the parameters to satisfy the m–condition, and search for the parameters which satisfy the e–condition (m–constraint algorithm).

It is easy to see that BP is an e–constraint algorithm since the e–condition is satisfied at each step, but its convergence is not necessarily guaranteed.

The other possibility is the m–constraint algorithm. We show that CCCP[14] is an m–constraint algorithm.

CCCP is an iterative double loop procedure to obtain the minimum of an energy function[13], and the idea was applied to the inference problem of loopy graphs, where the free energy in eq. (1) is the energy function [14]. The CCCP algorithm is defined as follows in information geometrical framework.

Information geometrical view of CCCP

inner loop: *Given* $\boldsymbol{\theta}^t$*, calculate* $\{\boldsymbol{\zeta}_r^{t+1}\}$ *by solving*

$$\pi_{M_0} \circ p_r(\boldsymbol{x}; \boldsymbol{\zeta}_r^{t+1}) = L\boldsymbol{\theta}^t - \sum_r \boldsymbol{\zeta}_r^{t+1}, \quad r = 1, \cdots, L. \tag{2}$$

outer loop: *Given a set of* $\{\boldsymbol{\zeta}_r^{t+1}\}$ *as the result of the inner loop, calculate*

$$\boldsymbol{\theta}^{t+1} = L\boldsymbol{\theta}^t - \sum_r \boldsymbol{\zeta}_r^{t+1}. \tag{3}$$

From eqs. (2) and (3), one obtains

$$\boldsymbol{\theta}^{t+1} = \pi_{M_0} \circ p_r(\boldsymbol{x}; \boldsymbol{\zeta}_r^{t+1}), \quad r = 1, \cdots, L,$$

which means that CCCP enforces the m–condition at each iteration. On the other hand, the e–condition is satisfied only at the convergent point, which can be easily verified by letting $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t = \boldsymbol{\theta}^*$ in eq. (3) to yield the e–condition $(L-1)\boldsymbol{\theta}^* = \sum_r \boldsymbol{\zeta}_r^*$. Therefore, we can see that the inner and outer loops of CCCP solve the m–condition and the e–condition, respectively.

The following theorem [7] shows CCCP has a good convergent property.

Theorem 5. *Let* K *be defined as follows*

$$K = \frac{1}{L}\sum_r \sqrt{I_0(\boldsymbol{\theta})} I_r(\boldsymbol{\zeta}_r)^{-1} \sqrt{I_0(\boldsymbol{\theta})}.$$

Let λ_i $i = 1, \cdots, n$ *be the eigen values of* K*, and the equilibrium of CCCP is stable when the eigen values of the following* K *satisfies*

$$\lambda_i > 1 - \frac{1}{L}, \quad i = 1, \cdots, n.$$

Under the m–constraint, the Hessian of $\mathcal{F}(\boldsymbol{\theta})$ at an equilibrium point is equal to

$$\sqrt{I_0(\boldsymbol{\theta})}\big[LK - (L-1)E\big]\sqrt{I_0(\boldsymbol{\theta})},$$

so that the stability condition for CCCP is equivalent to the condition that the equilibrium is a local minimum of $\mathcal{F}$ under the m–constraint.

3.2. **Adaptive TAP approximation.** We show that the adaptive TAP (Thouless Anderson Palmer) approximation is also formulated in a similar way by information geometry. We first summarize the results given by Opper and Winther[10]. Consider

$$q(\boldsymbol{x}) = \frac{1}{Z}\prod_{r=1}^n \rho(x_r) \exp\Big[\boldsymbol{h}\cdot\boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^T J \boldsymbol{x}\Big]. \tag{4}$$

J is a symmetric matrix where diagonal elements are 0. $\rho(x_r)$ can take a lot of kinds of functions, and in this extended memo, we consider the case $\rho(x_r)$ is strictly positive and integrable for $x_r \in \Re$.

The aim of the adaptive TAP approach is to infer $E_q[x_r]$ and $E_q[x_r^2]$. Let m_r be the inference of $E_q[x_r]$, and let us define $p_r(x_r)$ as follows,

$$p_r(x_r) = \frac{1}{Z_0^{(r)}}\rho(x_r)\exp\left[\Big(\sum_{s=1}^n J_{rs}m_s - V_r m_r + h_r\Big)x_r + \frac{V_r}{2}x_r^2\right], \tag{5}$$

where $Z_0^{(r)}$ is the normalization factor to make the integral of $p_r(x_r)$ equal to 1. The summary of adaptive TAP equations is given as follow.

$$m_r = \int x_r p_r(x_r) dx_r \tag{6}$$

$$\big[(S-J)^{-1}\big]_{rr} = \int (x_r - m_r)^2 p_r(x_r) dx_r = \int x_r^2 p_r(x_r) dx_r - m_r^2. \tag{7}$$

It is possible to view adaptive TAP equations as an variation of BP through information geometrical framework of BP. Here, we need to extend x_i from binary

variable to real variable. We consider the case where $\rho(x_r)$ is strictly positive for $x_r \in \Re$. Equation (4) can be rewritten as

$$q(\boldsymbol{x}) = \exp[c_0(\boldsymbol{x}) + c_1(\boldsymbol{x}) + \cdots + c_n(\boldsymbol{x}) - \psi_q],$$
$$c_0(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T J \boldsymbol{x}, \quad c_r(\boldsymbol{x}) = c_r(x_r) = \ln \rho(x_r) + h_r x_r, \quad \psi_q = \ln Z.$$

Next, we define p_0 as the distribution whose sufficient statistics are x_r, x_r^2, $r = 1, \cdots, n$. Natural choice is Gaussian distribution, which is defined as

$$p_0(\boldsymbol{x}; \boldsymbol{\mu}, S) = \exp\Big[c_0(\boldsymbol{x}) + \boldsymbol{\mu} \cdot \boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^T S \boldsymbol{x} - \psi_0(\boldsymbol{\mu}, S)\Big],$$
$$S = \mathrm{diag}(s_1, \cdots, s_n),\ \psi_0(\boldsymbol{\mu}, S) = \frac{n}{2}\ln 2\pi - \ln\det(S - J) + \frac{1}{2}\boldsymbol{\mu}^T (S - J)^{-1}\boldsymbol{\mu}.$$

We set M_0 as

$$M_0 = \{p_0(\boldsymbol{x}; \boldsymbol{\mu}, S) \mid \boldsymbol{\mu} \in \Re^n, S = \mathrm{diag}(s_1, \cdots, s_n), s_i > 0\}$$

Now, ultimate goal of our problem is to obtain the m–projection of $q(\boldsymbol{x})$ to M_0.

Let us define $p_r(\boldsymbol{x}; \boldsymbol{\mu}_{\backslash r}, S_{\backslash r})$ as follows

$$p_r(\boldsymbol{x}; \boldsymbol{\mu}_{\backslash r}, S_{\backslash r}) = \exp\Big[c_0(\boldsymbol{x}) + c_r(x_r) + \boldsymbol{\mu}_{\backslash r} \cdot \boldsymbol{x}_{\backslash r} - \frac{1}{2}{\boldsymbol{x}_{\backslash r}}^T S_{\backslash r} \boldsymbol{x}_{\backslash r} - \psi_r(\boldsymbol{\mu}_{\backslash r}, S_{\backslash r})\Big],$$
$$S_{\backslash r} = \mathrm{diag}(s_1, \cdots, s_{r-1}, s_{r+1}, \cdots, s_n) \in \Re^{(n-1)\times(n-1)},$$
$$\boldsymbol{\mu}_{\backslash r} = (\mu_1, \cdots, \mu_{r-1}, \mu_{r+1}, \cdots, \mu_n)^T \in \Re^{n-1},$$
$$\boldsymbol{x}_{\backslash r} = (x_1, \cdots, x_{r-1}, x_{r+1}, \cdots, x_n)^T \in \Re^{n-1}.$$

We define M_r as

$$M_r = \{p_r(\boldsymbol{x}; \boldsymbol{\mu}_{\backslash r}, S_{\backslash r})\}.$$

It is easy to show the e–condition holds,

$$q(\boldsymbol{x}) = \frac{1}{Z}\frac{\prod_{r=1}^n p_r(\boldsymbol{x}; \boldsymbol{\mu}_{\backslash r}, S_{\backslash r})}{p_0(\boldsymbol{x}; \boldsymbol{\mu}, S)^{n-1}}.$$

Moreover, we can show that the m–condition corresponds to the adaptive TAP equations.

Proposition 2. The m–condition is

$$p_0(\boldsymbol{x}; \boldsymbol{\mu}, S) = \pi_{M_0} \circ p_r(\boldsymbol{x}; \boldsymbol{\mu}_{\backslash r}, S_{\backslash r}), \quad r = 1, \cdots, n,$$

and it is satisfied, if and only if the following equations hold.

$$m_r = \int x_r p_r(x_r; \boldsymbol{\mu}_{\backslash r}, S_{\backslash r}) dx_r \tag{8}$$

$$[(S - J)^{-1}]_{rr} = \int x_r^2 p_r(x_r; \boldsymbol{\mu}_{\backslash r}, S_{\backslash r}) dx_r - m_r^2 \tag{9}$$

$$r = 1, \cdots, n, \quad \text{where} \quad \boldsymbol{m} = (S - J)^{-1}\boldsymbol{\mu}.$$

Now we move to the adaptive TAP equations

Lemma 1. *Equations* (8) *and* (9) *are equivalent to* (6) *and* (7).

Proof. Let us set $\boldsymbol{J}_r$, $K_{\backslash r}$, and $J_{\backslash r}$ as follows

$$J_{\backslash r} = \begin{pmatrix} 0 & \cdots & J_{1(r-1)} & J_{1(r+1)} & \cdots & J_{1n} \\ \vdots & \ddots & & & & \vdots \\ J_{(r-1)1} & & & & & J_{(r-1)n} \\ J_{(r+1)1} & & & & & J_{(r+1)n} \\ \vdots & & & & \ddots & \vdots \\ J_{n1} & \cdots & J_{n(r-1)} & J_{n(r+1)} & \cdots & 0 \end{pmatrix}$$

$$K_{\backslash r} = (S_{\backslash r} - J_{\backslash r}), \quad \boldsymbol{J}_r = (J_{r1}, \cdots, J_{r(r-1)}, J_{r(r+1)}, \cdots, J_{rn})^T$$

We can rewrite $p_0(\boldsymbol{x}; \boldsymbol{\mu}, S)$ and $p_r(\boldsymbol{x}; \boldsymbol{\mu}_{\backslash r}, S_{\backslash r})$ as follows,

$$p_0(\boldsymbol{x}; \boldsymbol{\mu}, S) = p_0(x_r; \boldsymbol{\mu}, S)\mathcal{N}({K_{\backslash r}}^{-1}(\boldsymbol{\mu}_{\backslash r} + \boldsymbol{J}_r x_r), {K_{\backslash r}}^{-1})$$

$$p_r(\boldsymbol{x}; \boldsymbol{\mu}_{\backslash r}, S_{\backslash r}) = p_r(x_r; \boldsymbol{\mu}_{\backslash r}, S_{\backslash r})\mathcal{N}({K_{\backslash r}}^{-1}(\boldsymbol{\mu}_{\backslash r} + \boldsymbol{J}_r x_r), {K_{\backslash r}}^{-1}) \tag{10}$$

It is easily proved that $p_r(x_r)$ in (5) is equivalent to $p_r(x_r; \boldsymbol{\mu}_{\backslash r}, S_{\backslash r})$ in (10), which proves the equivalence to the adaptive TAP equations. □

4. Conclusion

We have shown the summary of the information geometrical framework to analyze the BP algorithm[6, 7]. Since the idea of loopy BP is widely used in many fields, we hope further understanding of BP will be given from our framework. We have also shown that other types of propagation algorithms can be explained in this framework. This shows the generality of our information geometrical framework.

Acknowledgment

This work was supported by the Grant-in-Aid for Young Scientists (B), No. 16700227, MEXT, Japan.

References

[1] S. Amari and H. Nagaoka. *Methods of Information Geometry*. AMS and Oxford University Press, Providence, Rhode Island, 2000.

[2] C. Berrou, A. Glavieux, and P. Thitimajshima. Near Shannon limit error-correcting coding and decoding: Turbo-codes. In *Proceedings of IEEE International Conference on Communications*, pages 1064–1070, Geneva, Switzerland, May 1993.

[3] H. Bethe. Statistical theory of superlattices. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 150(871):552–575, July 1935.

[4] R. G. Gallager. Low density parity check codes. *IRE Transactions on Information Theory*, IT-8:21–28, January 1962.

[5] S. Ikeda, T. Tanaka, and S. Amari. Information geometrical framework for analyzing belief propagation decoder. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14, pages 407–414. MIT Press, Cambridge, MA, April 2002.

[6] S. Ikeda, T. Tanaka, and S. Amari. Information geometry of turbo and low-density parity-check codes. *IEEE Transactions on Information Theory*, 50(6):1097–1114, June 2004.

[7] S. Ikeda, T. Tanaka, and S. Amari. Stochastic reasoning, free energy, and information geometry. *Neural Computation*, 16(9):1779–1810, September 2004.

[8] M. I. Jordan. *Learning in Graphical Models*. MIT Press, Cambridge, MA, 1999.

[9] S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society B*, 50:157–224, 1988.

[10] M. Opper and O. Winther. Adaptive and self-averaging Thouless-Anderson-Palmer mean field theory for probabilistic modeling. *Physical Review E*, 64:056131, 2001.

[11] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo. CA, 1988.

[12] Y. Weiss. Correctness of local probability propagation in graphical models with loops. *Neural Computation*, 12(1):1–41, January 2000.

[13] A. L. Yuille. CCCP algorithms to minimize the Bethe and Kikuchi free energies: Convergent alternatives to belief propagation. *Neural Computation*, 14(7):1691–1722, July 2002.
[14] A. L. Yuille and A. Rangarajan. The concave-convex procedure. *Neural Computation*, 15(4):915–936, April 2003.

The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo, 106-8569, Japan
E-mail address: shiro@ism.ac.jp
URL: http://www.ism.ac.jp/~shiro