Information Geometry of Loopy BP

Shiro Ikeda
Inst. of Statistical Mathematics
shiro@ism.ac.jp

Toshiyuki Tanaka
Tokyo Metropolitan Univ.
tanaka@eei.metro-u.ac.jp

Shun-ichi Amari
RIKEN BSI
amari@brain.riken.go.jp

Abstract

Belief propagation (BP) is an efficient algorithm to solve the inference problem of graphical models. We give the information geometrical view of the algorithm, and propose a new cost function which yields a new algorithm.

1 Introduction

Although Pearl's belief propagation algorithm[3] was only proved to give the exact inference for tree graphs, a lot of applications suggest it also works well for loopy graphs.

We have given the information geometrical framework [1] to analyze the BP decoding algorithms of turbo codes and LDPC codes[2, 4]. In this article, we extend our results to the BP algorithm on general graphs. The main results are the characterization of the fixed points and geometrical explanation about the cause of error of the inference. We also gives a new cost function which yields a new algorithm.

2 Belief Propagation

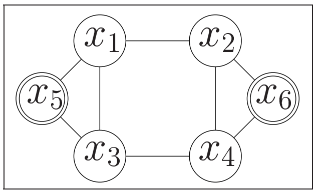

Figure 1: An example of undirected graphs: each circle shows a node, and double circles are the observed nodes.

We discuss the inference of undirected graphs in this article. Let $\{x_1, \cdots, x_n\}$ be the set of nodes and $\mathcal{L}$ be that of links. Each x_i is a binary stochastic variable $x_i \in \{-1, +1\}$, and we set first m nodes to be hidden and the rest to be observed. In Fig.1, $m = 4$, $n = 6$, $\mathcal{L} = \{(1,2),(1,3),(1,5),(2,4),(2,6),(3,4),(3,5),(4,6)\}$.

We consider only the binary stochastic variables in this article, but the results of this article can be easily generalized to any discrete stochastic variables.

The joint probability distribution of $\boldsymbol{x}$ is given as follows,

$$q(\boldsymbol{x}) = \frac{1}{Z} \prod_{(i,j)\in\mathcal{L}} \psi_{ij}(x_i, x_j),$$

where Z is the normalization factor and $\psi_{ij}(x_i, x_j) > 0$. Let $\boldsymbol{y} = (x_1, \cdots, x_m)^T$ denote hidden variables, and $\boldsymbol{z} = (x_{m+1}, \cdots, x_n)^T$ denote observed nodes (^{T} represents transpose), and let us consider $q(\boldsymbol{y}|\boldsymbol{z})$,

$$q(\boldsymbol{y}) \overset{\text{def}}{=} q(\boldsymbol{y}|\boldsymbol{z}) = \frac{1}{Z} \prod_{i=1}^{m} \varpi_i(y_i) \prod_{(i,j)\in\mathcal{L}:i,j\le m} \psi_{ij}(y_i, y_j), \quad (1)$$

where $\varpi_i(y_i)$ is defined as follows,

$$\varpi_i(y_i) \overset{\text{def}}{=} \prod_{j:(i,j)\in\mathcal{L}:i\le m, j>m} \psi_{ij}(y_i, z_{j-m}).$$

The goal of the BP algorithm is to infer the marginal distribution, $q(y_i)$, of $q(\boldsymbol{y})$. We infer $q(y_i)$ as $b_i(y_i)$. Generally, the marginalization of $q(\boldsymbol{y})$ is NP-hard, but for tree graphs, the BP algorithm is efficient and gives the exact inference. The original BP algorithm is given below following the notations of Yedidia et.al[6].

BP algorithm

1. Set $t = 0$, and $m_{ij}^t(y_j) = 1/2$, for $\forall(i,j) \in \mathcal{L}$.
2. For $t = 1, 2, \cdots$, update messages as follows,

$$m_{ij}^{t+1}(y_j) = \frac{1}{Z} \sum_{y_i} \varpi_i(y_i) \psi_{ij}(y_i, y_j) \prod_{k\in\mathcal{N}(i)\setminus j} m_{ki}^t(y_i), \quad (2)$$

   here Z normalizes as $\sum_{y_j} m_{ij}^{t+1}(y_j) = 1$, $\mathcal{N}(i)$ is the set of nodes connected to node i. Belief is given as

$$b_i(y_i) = \frac{1}{Z} \varpi_i(y_i) \prod_{k\in\mathcal{N}(i)} m_{ki}^{t+1}(y_i). \quad (3)$$

3. Repeat step 2 until $b_i(y_i)$ converges.

It is well-known that, for loopy graphs, the BP algorithm does not necessarily converge, and $b_i(y_i)$ is not exactly equal to $q(y_i)$ even if it converges.

3 Preliminaries of Information Geometry

We give the preliminaries of information geometry.

Since every multinomial distribution is an exponential family[1], the set of all the probability distributions on $\boldsymbol{y}$ is an exponential family.

$$S \overset{\text{def}}{=} \Big\{ p(\boldsymbol{y}) | p(\boldsymbol{y}) > 0, \sum_{\boldsymbol{y}} p(\boldsymbol{y}) = 1 \Big\}.$$

Next, we define e–flat and m–flat submanifolds of S.

e–**flat submanifold:** $M{\subset}S$ is e–flat, when $r(\boldsymbol{y};s)$ belongs to M for all $q(\boldsymbol{y}), p(\boldsymbol{y}) \in M$,

$$\ln r(\boldsymbol{y};s) = (1-s)\ln q(\boldsymbol{y}) + s\ \ln p(\boldsymbol{y}) + c(s), s{\in}R,$$

where $c(s)$ is the normalization factor.

m–**flat submanifold:** $M{\subset}S$ is m–flat, when $r(\boldsymbol{y};s)$ belongs to M for all $q(\boldsymbol{y}), p(\boldsymbol{y}) \in M$,

$$r(\boldsymbol{y};s) = (1-s)q(\boldsymbol{y}) + sp(\boldsymbol{y}), \qquad s{\in}[0,1].$$

Next, we define the m–projection.

Definition 1. *Let $M{\subset}S$, and $q(\boldsymbol{y}){\in}S$. The point in M that minimizes the KL divergence from $q(\boldsymbol{y})$ to M,*

$$p^*(\boldsymbol{y}) = \underset{p(\boldsymbol{y})\in M}{\operatorname{argmin}} D[q(\boldsymbol{y});p(\boldsymbol{y})], \tag{4}$$

is called the m–projection of $q(\boldsymbol{y})$ to M.

The m–projection theorem follows[1].

Theorem 1. *Let M be an e–flat submanifold in S, and let $q(\boldsymbol{y}){\in}S$. The m–projection of $q(\boldsymbol{y})$ to M is unique.*

The KL divergence, $D[\cdot;\cdot]$ is defined as follows,

$$D\left[q(\boldsymbol{y});p(\boldsymbol{y})\right] = \sum_{\boldsymbol{y}} q(\boldsymbol{y}) \ln \frac{q(\boldsymbol{y})}{p(\boldsymbol{y})} \geq 0,$$

if $q(\boldsymbol{y}) = p(\boldsymbol{y})$ holds for every $\boldsymbol{y}$, $D[q(\boldsymbol{y});p(\boldsymbol{y})] = 0$.

4 Information Geometrical View of BP

4.1 Information Geometrical View of Belief

Since each $y_i \in \{-1,+1\}$, we can define $b_i(y_i)$ as,

$$b_i(y_i) \overset{\text{def}}{=} b_i(y_i;\theta^i) {=} \exp(k_i(y_i) + \theta^i y_i - \phi_i(\theta^i)), \tag{5}$$

where $\theta^i{\in}\mathcal{R}$ is the natural parameter[1] and $\phi_i(\theta^i)$ is the normalization factor. From eq.(3), we set $k_i(y_i) = \ln \varpi_i(y_i)$. Let us consider the distribution of $\boldsymbol{y}$, where the distribution of each y_i is $b_i(y_i;\theta^i)$ and is independent.

$$p_0(\boldsymbol{y};\boldsymbol{\theta}) \overset{\text{def}}{=} \prod_{i=1}^m b_i(y_i;\theta^i) = \exp\bigl(k_0(\boldsymbol{y}) + \boldsymbol{\theta}\cdot\boldsymbol{y} - \varphi_0(\boldsymbol{\theta})\bigr)$$

$$k_0(\boldsymbol{y}){=}\sum_{i=1}^m k_i(y_i), \varphi_0(\boldsymbol{\theta}){=}\sum_{i=1}^m \phi_i(\theta^i), \boldsymbol{\theta}{=}(\theta^1,\cdots,\theta^m)^T.$$

We define the manifold of $p_0(\boldsymbol{y};\boldsymbol{\theta})$ as

$$M_0{=}\Bigl\{p_0(\boldsymbol{y};\boldsymbol{\theta}){=}\exp\bigl(k_0(\boldsymbol{y}) + \boldsymbol{\theta}{\cdot}\boldsymbol{y} - \varphi_0(\boldsymbol{\theta})\bigr)\big|\boldsymbol{\theta} \in \mathcal{R}^m\Bigr\}.$$

M_0 is an e–flat submanifold. Next, we define another coordinate system $\boldsymbol{\eta}_0 = (\eta_{01},\cdots,\eta_{0m})^T$ called the expectation parameter[1]

$$\boldsymbol{\eta}_0 = E_{p_0(\boldsymbol{y};\boldsymbol{\theta})}[\boldsymbol{y}] = \partial_{\boldsymbol{\theta}}\varphi_0(\boldsymbol{\theta}),\ \eta_{0i} = \sum_{y_i} b_i(y_i;\theta^i)y_i,$$

where $E_p[\cdot]$ is the expectation with respect to p. We denote $\boldsymbol{\eta}_0$ as $\boldsymbol{\eta}_0(\boldsymbol{\theta})$ for the following of the article, since $\boldsymbol{\eta}_0$ is a function of $\boldsymbol{\theta}$. The ideal goal of the BP algorithm is to infer a point in M_0 which corresponds to the product of marginal distributions, that is, $\prod_i q(y_i)$. Now, we redefine $q(\boldsymbol{y})$ in eq.(1) as

$$q(\boldsymbol{y}) = \frac{1}{Z}\exp\Bigl(k_0(\boldsymbol{y}) + \sum_{(i,j)\in\mathcal{L}} c_{ij}(\boldsymbol{y})\Bigr). \tag{6}$$

where $c_{ij}(\boldsymbol{y}) = \ln\psi_{ij}(y_i,y_j)$.

Proposition 1. *Marginalized distribution of $q(\boldsymbol{y})$, that is, $\prod_i q(y_i)$, is the m–projected point from $q(\boldsymbol{y})$ to M_0.*

Proof. The m–projection from $q(\boldsymbol{y})$ to M_0 minimizes the KL divergence $D[q(\boldsymbol{y});p_0(\boldsymbol{y};\boldsymbol{\theta})]$. Since M_0 is e–flat, the point is unique. The derivative of the KL divergence with respect to $\boldsymbol{\theta}$ is,

$$\partial_{\boldsymbol{\theta}} D[q(\boldsymbol{y});p_0(\boldsymbol{y};\boldsymbol{\theta})] = \boldsymbol{\eta}_0(\boldsymbol{\theta}) - E_{q(\boldsymbol{y})}[\boldsymbol{y}].$$

As the stationary condition we have, $\boldsymbol{\eta}_0(\boldsymbol{\theta}) = E_{q(\boldsymbol{y})}[\boldsymbol{y}]$. The m–projection of $q(\boldsymbol{y})$ to M_0 does not change the expectation of $\boldsymbol{y}$ and equivalent to the marginalization of $q(\boldsymbol{y})$. We define the operator π_{M_0} which gives the m–projected parameter as

$$\pi_{M_0} \circ q(\boldsymbol{y}) = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} D[q(\boldsymbol{y});p_0(\boldsymbol{y};\boldsymbol{\theta})]. \quad \square$$

4.2 Messages and Links

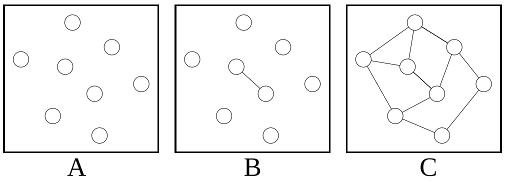

Figure 2: A: Belief graph, B: Graph with a single link, C: Graph with all the links.

Figure 2 shows 3 important graphs in BP. The belief graph in Fig.2.A corresponds to $p_0(\boldsymbol{y};\boldsymbol{\theta})$, and next, we consider the distribution $p_r(\boldsymbol{y})$, which includes a single link $\psi_{ij}(y_i,y_j)$ (Fig.2.B). We start with reconsidering eq.(3) in terms of θ^i of belief (eq.(5))

$$\begin{aligned} b_i(y_i;\theta^i) &= \frac{1}{Z}\varpi_i(y_i)\prod_{k\in\mathcal{N}(i)} m_{ki}(y_i) \\ &= \exp\bigl(k_i(y_i) + \theta^i\cdot y_i - \phi_i(\theta^i)\bigr), \end{aligned}$$

from the fact $k_i(y_i) = \ln\varpi_i(y_i)$, the rest of the problem is the relation between $m_{ki}(y_i)$ and θ^i. We introduce μ^i_k as

$$\frac{m_{ki}(y_i)}{\sum_{y_i} m_{ki}(y_i)} = \frac{1}{2}(\tanh(\mu^i_k y_i) + 1),$$

$$b_i(y_i;\theta^i) = \exp\Bigl(k_i(y_i) {+} \sum_{k\in\mathcal{N}(i)} \mu^i_k y_i {-} \phi_i\bigl(\sum_{k\in\mathcal{N}(i)} \mu^i_k\bigr)\Bigr).$$

$m_{ki}(y_i)$ and μ_k^i have one to one relation, and θ^i is rewritten as $\theta^i = \sum_{k \in \mathcal{N}(i)} \mu_k^i$. In order to make the following discussion clear, we redefine the $\{\mu_k^i\}$ as,

$$\boldsymbol{\xi}_r{=}(0,\cdots,0,\underset{\hat{i}}{\mu_j^i},0,\cdots,0,\underset{\hat{j}}{\mu_i^j},0,\cdots,0)^T, \boldsymbol{\theta} = \sum_{r \in \mathcal{L}} \boldsymbol{\xi}_r.$$

Here, r is the new index to indicate (i,j). Since $p_r(\boldsymbol{y})$ includes $\psi_{ij}(y_i, y_j)$, the messages $m_{ij}(y_j)$ and $m_{ji}(y_i)$ should be eliminated. The parameter of $p_r(\boldsymbol{y})$ is $\boldsymbol{\zeta}_r = \boldsymbol{\theta} - \boldsymbol{\xi}_r$, and $p_r(\boldsymbol{y}; \boldsymbol{\zeta}_r)$ is defined as,

$$p_r(\boldsymbol{y};\boldsymbol{\zeta}_r){=}\exp(k_0(\boldsymbol{y}) + c_r(\boldsymbol{y}) + \boldsymbol{\zeta}_r \cdot \boldsymbol{y} - \varphi_r(\boldsymbol{\zeta}_r)),$$
$$c_r(\boldsymbol{y}) \overset{\text{def}}{=} \ln \psi_{ij}(y_i, y_j).$$

The definitions of an e–flat submanifold M_r, which is the family of $p_r(\boldsymbol{y}; \boldsymbol{\zeta}_r)$, and of the expectation parameter $\boldsymbol{\eta}_r(\boldsymbol{\zeta}_r)$ are given as follows for $r \in \mathcal{L}$,

$$M_r = \Big\{ p_r(\boldsymbol{y}; \boldsymbol{\zeta}_r) \Bigm| \boldsymbol{\zeta}_r {\in} \mathcal{R}^m \Big\},$$
$$\boldsymbol{\eta}_r(\boldsymbol{\zeta}_r) = \sum_{\boldsymbol{y}} p_r(\boldsymbol{y}; \boldsymbol{\zeta}_r)\boldsymbol{y} = \partial_{\boldsymbol{\zeta}_r} \varphi_r(\boldsymbol{\zeta}_r).$$

Now, we show the information geometrical view of the BP algorithm. Let us reconsider eq.(2) by multiplying $\varpi_j(y_j) \prod_{k \in \mathcal{N}(j) \backslash i} m_{kj}^t(y_j)$ on both sides of eq.(2). The right hand side is rewritten as,

$$\frac{1}{Z} \sum_{y_i} \varpi_i(y_i) \varpi_j(y_j) \psi_{ij}(y_i, y_j) \prod_{k \in \mathcal{N}(j) \backslash i} m_{kj}^t(y_j) \prod_{k \in \mathcal{N}(i) \backslash j} m_{ki}^t(y_i)$$
$$= \sum_{y_k : k \neq j} p_r(\boldsymbol{y}; \boldsymbol{\zeta}_r^t),$$

while the left hand side is,

$$\frac{1}{Z} \varpi_j(y_j) m_{ij}^{t+1}(y_j) \prod_{k \in \mathcal{N}(j) \backslash i} m_{kj}^t(y_j)$$
$$= \sum_{y_k : k \neq j} p_0(\boldsymbol{y}; \boldsymbol{\theta}^t - \boldsymbol{\xi}_r^t + \boldsymbol{\xi}_r^{t+1}) {=} \sum_{y_k : k \neq j} p_0(\boldsymbol{y}; \boldsymbol{\zeta}_r^t + \boldsymbol{\xi}_r^{t+1}).$$

From Proposition 1 and by assuming the case $m_{ij}(y_j)$ and $m_{ji}(y_i)$ are updated simultaneously, eq.(2) is rewritten as follows,

$$\boldsymbol{\zeta}_r^t + \boldsymbol{\xi}_r^{t+1} = \pi_{M_0} \circ p_r(\boldsymbol{y}; \boldsymbol{\zeta}_r^t).$$

In summary, the BP algorithm is expressed as follows in terms of information geometry.

Information geometrical view of the BP algorithm

1. Set $t = 0$, $\boldsymbol{\theta}^t = \mathbf{o}$, $\boldsymbol{\zeta}_r^t = \boldsymbol{\xi}_r^t = \mathbf{o}$, $r \in \mathcal{L}$.
2. For $t = 1, 2, \cdots$, update $\boldsymbol{\xi}_r^{t+1}$, as follows,

   $$\boldsymbol{\xi}_r^{t+1} = \pi_{M_0} \circ p_r(\boldsymbol{y}; \boldsymbol{\zeta}_r^t) - \boldsymbol{\zeta}_r^t.$$

   $\boldsymbol{\theta}^{t+1} = \sum_{r \in \mathcal{L}} \boldsymbol{\xi}_r^{t+1}$ and $\boldsymbol{\zeta}_r^{t+1} = \boldsymbol{\theta}^{t+1} - \boldsymbol{\xi}_r^{t+1}$.
3. If $\boldsymbol{\xi}_r^{t+1}$ does not converge, $t{+}1 {\rightarrow} t$ and go to 2.

4.3 Fixed Points of BP

The fixed points of the BP algorithm satisfy the following conditions,

$$1) \quad \boldsymbol{\eta}_0(\boldsymbol{\theta}^*) = \boldsymbol{\eta}_r(\boldsymbol{\zeta}_r^*), r \in \mathcal{L} \tag{7}$$

$$2) \quad \boldsymbol{\theta}^* = \sum_{r \in \mathcal{L}} \boldsymbol{\xi}_r^* = \frac{1}{L-1} \sum_{r \in \mathcal{L}} \boldsymbol{\zeta}_r^*, \tag{8}$$

where * denotes the fixed point and L is the number of the links. Let us define two submanifolds, one is an m–flat submanifold M^*,

$$M^* = \Big\{ p(\boldsymbol{y}) \Bigm| \sum_{\boldsymbol{y}} p(\boldsymbol{y}) \boldsymbol{y} = \boldsymbol{\eta}_0(\boldsymbol{\theta}^*) \Big\}.$$

Expectation of $\boldsymbol{y}$ is the same for every $p(\boldsymbol{y}) {\in} M^*$, and m–projection from $p(\boldsymbol{y}) {\in} M^*$ to M_0 coincides with $p_0(\boldsymbol{y}; \boldsymbol{\theta}^*)$. The other is an e–flat submanifold E^*

$$E^* {=} \Big\{ p(\boldsymbol{y}) {=} \frac{1}{Z} p_0(\boldsymbol{\theta}^*)^{t_0} \prod_{r \in \mathcal{L}} p_r(\boldsymbol{\zeta}_r^*)^{t_r} \Big| t_* {\in} \mathcal{R}, t_0 {+} \sum_{r \in \mathcal{L}} t_r {=} 1 \Big\},$$

The distributions $p_0(\boldsymbol{y}; \boldsymbol{\theta}^*)$, $p_r(\boldsymbol{y}; \boldsymbol{\zeta}_r^*)$ $(r \in \mathcal{L})$ are included in M^*. Moreover, we can check that $q(\boldsymbol{y})$ is included in E^*. This result leads us to the following theorem.

Theorem 2. *At the fixed points of the BP algorithm, $p_0(\boldsymbol{y}; \boldsymbol{\theta}^*)$ and $p_r(\boldsymbol{y}; \boldsymbol{\zeta}_r^*)$, ($r \in \mathcal{L}$) are included in the m–flat submanifold M^*, while the e–flat submanifold E^* includes $p_0(\boldsymbol{y}; \boldsymbol{\theta}^*)$, $p_r(\boldsymbol{y}; \boldsymbol{\zeta}^*)$, ($r \in \mathcal{L}$), and $q(\boldsymbol{y})$.*

If M^* includes $q(\boldsymbol{y})$, $p_0(\boldsymbol{y}; \boldsymbol{\theta}^*)$ gives the true belief, but $q(\boldsymbol{y})$ is only included in E^*, and generally there is a discrepancy between M^* and E^* for loopy graphs. This discrepancy gives the difference between the true and the inferred beliefs.

5 New Cost Function

The fixed point of the BP algorithm is characterized with eqs.(7) and (8).

Some BP related algorithms, such as CCCP[7], obtain the fixed point by double loops algorithms. The inner loop adjusts parameters to satisfy eq.(7), and the outer loop adjusts them to satisfy eq.(8). This is one idea to have a general convergence, but we consider another possibility. We constrain the parameters to satisfy eq.(8), and search for the parameters which satisfy eq.(7). In order to make this possible, we start with proposing a new cost function as

$$F(\{\boldsymbol{\zeta}_r\}) {=} \sum_{r \in \mathcal{L}} ||\boldsymbol{\eta}_0(\boldsymbol{\theta}) {-} \boldsymbol{\eta}_r(\boldsymbol{\zeta}_r)||^2.$$

If the cost function is minimized under the constraint, $\boldsymbol{\theta} = \sum_{r \in \mathcal{L}} \boldsymbol{\zeta}_r / (L-1)$, both of eq.(7) and eq.(8) are satisfied at the minimum point, where $F = 0$. A naive method to minimize F is the gradient descent. Under the constraint of eq.(8), the gradient is

$$\begin{aligned}\frac{\partial F}{\partial \boldsymbol{\zeta}_r} =& -2I_r(\boldsymbol{\zeta}_r)(\boldsymbol{\eta}_0(\boldsymbol{\theta})-\boldsymbol{\eta}_r(\boldsymbol{\zeta}_r)) \\ &+ \frac{2}{L-1} I_0(\boldsymbol{\theta}) \sum_r (\boldsymbol{\eta}_0(\boldsymbol{\theta})-\boldsymbol{\eta}_r(\boldsymbol{\zeta}_r)).\end{aligned} \tag{9}$$

Here, $I_0(\boldsymbol{\theta})$ and $I_r(\boldsymbol{\zeta}_r)$ are the Fisher information matrices of $p_0(\boldsymbol{y};\boldsymbol{\theta})$ and $p_r(\boldsymbol{y};\boldsymbol{\zeta}_r)$, respectively. If the derivative is available, $\boldsymbol{\zeta}_r$ and $\boldsymbol{\theta}$ are updated as,

$$\boldsymbol{\zeta}_r^{t+1} = \boldsymbol{\zeta}_r^t - \delta \frac{\partial F}{\partial \boldsymbol{\zeta}_r^t}, \quad \boldsymbol{\theta}^{t+1} = \frac{1}{L} \sum_{r \in \mathcal{L}} \boldsymbol{\zeta}_r^{t+1}.$$

where δ is a small positive learning rate. Since $p_0(\boldsymbol{y};\boldsymbol{\theta})$ is factorisable distribution, it is easy to calculate $\boldsymbol{\eta}_0(\boldsymbol{\theta})$ from $\boldsymbol{\eta}_0(\boldsymbol{\theta}) = \sum_{\boldsymbol{y}} p_0(\boldsymbol{y};\boldsymbol{\theta})\boldsymbol{y}$. Also $I_0(\boldsymbol{\theta})$ is simply calculated as,

$$I_0(\boldsymbol{\theta}) = \sum_{\boldsymbol{y}} p_0(\boldsymbol{y};\boldsymbol{\theta})(\boldsymbol{y}-\boldsymbol{\eta}_0(\boldsymbol{\theta}))(\boldsymbol{y}-\boldsymbol{\eta}_0(\boldsymbol{\theta}))^T.$$

With the BP algorithm, $\pi_{M_0} \circ p_r(\boldsymbol{y};\boldsymbol{\zeta}_r^t)$ is tractable, and $\boldsymbol{\eta}_r(\boldsymbol{\zeta}_r)$ is calculated from the relation,

$$\boldsymbol{\eta}_r(\boldsymbol{\zeta}_r) = \boldsymbol{\eta}_0(\pi_{M_0} \circ p_r(\boldsymbol{y};\boldsymbol{\zeta}_r^t)).$$

We have shown the calculations of $\boldsymbol{\eta}_0(\boldsymbol{\theta})$, $\boldsymbol{\eta}_r(\boldsymbol{\zeta}_r)$, and $I_0(\boldsymbol{\theta})$ are tractable. The rest of the problem is to calculate the first term of eq.(9). Fortunately, we have the relation,

$$I_r(\boldsymbol{\zeta}_r)\boldsymbol{h} = \lim_{\alpha \to 0} \frac{\boldsymbol{\eta}_r(\boldsymbol{\zeta}_r + \alpha \boldsymbol{h}) - \boldsymbol{\eta}_r(\boldsymbol{\zeta}_r)}{\alpha}.$$

If $\boldsymbol{h}$ is substituted with $(\boldsymbol{\eta}_0(\boldsymbol{\theta})-\boldsymbol{\eta}_r(\boldsymbol{\zeta}_r))$, this becomes the first term of eq.(9). Now, we propose a new algorithm.

New algorithm

1. Set $t = 0$, $\boldsymbol{\theta}^t = \mathbf{o}$, $\boldsymbol{\zeta}_r^t = \mathbf{o}$, $r \in \mathcal{L}$.
2. Calculate $\boldsymbol{\eta}_0(\boldsymbol{\theta})$, $I_0(\boldsymbol{\theta})$, and $\boldsymbol{\eta}_r(\boldsymbol{\zeta}_r)$, $r \in \mathcal{L}$ with BP.
3. Let $\boldsymbol{h}_r {=} \boldsymbol{\eta}_0(\boldsymbol{\theta}) {-} \boldsymbol{\eta}_r(\boldsymbol{\zeta}_r)$ and calculate $\boldsymbol{\eta}_r(\boldsymbol{\zeta}_r {+} \alpha \boldsymbol{h}_r)$ for $r {\in} \mathcal{L}$, where $\alpha {>} 0$ is small. Then calculate
$$\boldsymbol{g}_r = \frac{\boldsymbol{\eta}_r(\boldsymbol{\zeta}_r + \alpha \boldsymbol{h}) - \boldsymbol{\eta}_r(\boldsymbol{\zeta}_r)}{\alpha}.$$
4. For $t = 1, 2, \cdots$, update $\boldsymbol{\zeta}_r^{t+1}$ as follows,
$$\boldsymbol{\zeta}_r^{t+1} = \boldsymbol{\zeta}_r^t - \delta\big(-2\boldsymbol{g}_r + \frac{2}{L-1} I_0(\boldsymbol{\theta}) \sum_r \boldsymbol{h}_r\big)$$
$$\boldsymbol{\theta}^{t+1} = \sum_{r \in \mathcal{L}} \boldsymbol{\zeta}_r^{t+1}/(L-1).$$
5. If $F(\{\boldsymbol{\zeta}_r\}) {=} \sum_{r \in \mathcal{L}} ||\boldsymbol{\eta}_0(\boldsymbol{\theta}) {-} \boldsymbol{\eta}_r(\boldsymbol{\zeta}_r)||^2 > \epsilon$ (ϵ is the threshold) holds, $t{+}1 {\to} t$ and go to 2.

This algorithm does not include double loops, which is different from CCCP, but includes new adjustable parameters δ and α. The choice of them is one of our future works. Another issue is the minimization techniques. It is possible to apply quasi-Newton methods.

6 Conclusion and Future Work

We have shown an information geometrical framework to understand and to analyze the BP algorithm in this article. The information geometrical structure of the fixed point is summarized in Theorem 2. It shows that the e–flat submanifold E^* and the m–flat submanifold M^* play an important role for the BP algorithm. The conditions of the BP fixed points are summarized in eq.(7) and eq.(8). Recently, many BP related algorithms are proposed[5, 7], and information geometrical will help to give a uniform view of them.

We also proposed a new variant with a new cost function. In this community, the Bethé free energy is a well-known cost function[6]. It has been shown the Bethé free energy is deeply related to the BP algorithm, but the property of it is not well understood. In Section 5, we proposed a new cost function. It is clearly shown that the cost function is 0 at the fixed points of the BP algorithm, and it is the minimum. We have shown the gradient descent algorithm for minimizing the new cost function, which does not have double loops. The BP algorithm is used twice to calculate the gradient. There are a lot of possible extensions in this direction. We can consider similar quadratic cost functions with different measures, and can apply other minimization techniques. These are a part of future works.

This paper gives a first step to the information geometrical understanding of the BP algorithm. We believe further study in this direction will lead us to better understanding and improvements of the BP algorithm.

References

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

[2] S. Ikeda, T. Tanaka, and S. Amari. Information geometrical framework for analyzing belief propagation decoder. In *NIPS 14*, The MIT Press, 2002.

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

[4] T. Tanaka, S. Ikeda, and S. Amari. Information-geometrical significance of sparsity in Gallager codes. In *NIPS 14*, The MIT Press, 2002.

[5] Y. W. Teh and M. Welling. The unified propagation and scaling algorithm. In *NIPS 14*, The MIT Press, 2002.

[6] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Bethe free energy, Kikuchi approximations, and belief propagation algorithms. MERL TR2001–16, 2001.

[7] A. L. Yuille and A. Rangarajan. The concave-convex procedure (CCCP). In *NIPS 14*, The MIT Press, 2002.