BN2002 ベイジアンネットセミナー 2002
Seminer on Bayesian Networks (BN2002)
Tokyo, Japan, Sep 1-2, 2002.

確率伝搬法の情報幾何

池田 思朗[*] 田中 利幸[†] 甘利 俊一[‡]
Shiro Ikeda Toshiyuki Tanaka Shun-ichi Amari

Abstract: 本稿では，グラフィカルモデルにおける確率伝搬法を情報幾何学によって表現した結果を示す．本研究の目的は，確率伝搬法の解析を行なうための見通しが良い数学的枠組を与えることである．本稿の結果を発展させることで，確率伝搬法の様々な性質が明らかになると考える．なお，本稿の結果は有向，無向を含む全てのグラフィカルモデルに対して成り立つ．したがって，ベイジアンネット，誤り訂正符号，Bethé 近似，クラスタ変分法を含む広いモデルに対して有効である．

1 はじめに

グラフィカルモデルの特徴は，全ての確率変数の同時確率密度関数を局所的な確率変数間の関係として表現することである．その結果，視覚的にも数式的にも直感的に分り易い確率分布の表現を与え，広いクラスの確率分布を表現できる．しかし，表現は簡単であるにもかかわらず，困難な問題が数多く残されている．1 つの大きな問題はデータからのモデルの推定法である．グラフィカルモデルのモデル選択，パラメータ推定は大きな問題であり，未だに困難な問題が残っている．一方，構造とパラメータが既定された上で，各確率変数に関する周辺分布，すなわちビリーフを計算する手法も実用上大きな問題である．本稿ではこのビリーフの計算法に焦点を当てる．

グラフの規模が小さい場合には，ノードの取り得る状態を全て計算すればビリーフを正しく計算できる．全てのノードが 2 値の確率変数であるとすれば，この計算のためにはノード数 N に対して 2^N 個の項の和が必要になる．したがって，グラフの規模が大きい場合には計算量が発散してしまい，このような方法ではビリーフの計算はできない．一方，グラフの規模が大きい場合も，グラフが “木” の構造をしているならば，確率伝搬法と呼ばれる手法により有限回の計算でビリーフが正しく求まることが分っており，理論的には解決している [9]．しかし，グラフの構造にループがある場合には，一部のグラフに対する実験的そして理論的結果があるものの，一般のグラフに対しては未解決の問題が多く残っている．

ループのあるグラフィカルモデルのビリーフの計算は，応用上重要な分野を含んでいる．1 つの重要な分野は誤り訂正符号である．ターボ符号，そして低密度パリティ検査 (LDPC) 符号と呼ばれる誤り訂正符号では，確率伝搬法に相当する手法を復号法として用いることで大きな成功を納めている [7, 8]．また，統計物理の近似手法である Bethé 近似，クラスタ変分法も確率伝搬法として理解できることが分っており [4]，その有効性も実験的に確かめられている．

我々は，情報幾何 [2] によって誤り訂正符号の解析のための数学的枠組を与え，誤り訂正符号の繰り返し復号法の持つ性質を明らかにした [3, 11]．本論文ではその結果を拡張し，一般のグラフィカルモデルの確率伝搬法を解析するための枠組を与える．

2 グラフィカルモデルと確率伝搬法

2.1 無向グラフの表現

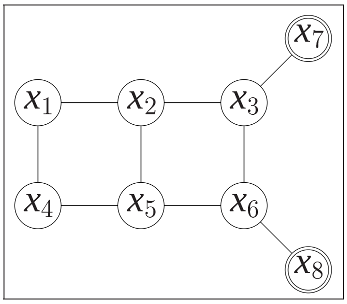


図 1: グラフィカルモデルの例

図 1 にグラフィカルモデルの例を挙げる．本稿では無

*九工大 & 科技団, 〒 808-0196 福岡県 北九州市 若松区 ひびきの 2–4, Tel. 093-695-6124, e-mail shiro@brain.kyutech.ac.jp,
Kyushu Institute of Technology & JST, 2–4 Hibikino, Wakamatsu, Kitakyushu, Fukuoka 8080196, Japan.
†東京都立大学, 〒 192-0397 東京都 八王子市 南大沢 1–1, e-mail tanaka@eei.metro-u.ac.jp,
Tokyo Metropolitan University, 1–1 Minami Oosawa, Hachioji, Tokyo, 1920397 Japan.
‡理研 脳総研, 〒 351-0198 埼玉県 和光市 広沢 2–1,
RIKEN, BSI, 2–1 Hirosawa, Wako, Saitama, 3510198, Japan.

向グラフのみを扱う．$\{x_1, \cdots, x_n\}$ をノードとする．これらは確率変数である．本稿では各ノードは 2 値の確率変数であり，$x_i \in \{-1, +1\}$ とする．任意の離散変数への拡張は簡単である．

一般的な確率伝搬法の問題では，一部のノードが観測されている場合を扱う．図中に 2 重で示したノードは，観測されているとする．なお，議論を簡単にするため，観測されていないノードの数を m とし x_i の添字 i は，$1 \leq i \leq m$ ならば非観測ノード，$m+1 \leq i \leq n$ ならば観測ノードであるように順番がつけられているとする．これによって議論の一般性が損なわれることはない．上の例であれば，$n = 8$, $m = 6$, である．非観測ノードからなるベクトルを $\boldsymbol{y}$, 観測ノードのみからなるベクトルを $\boldsymbol{z}$ とし，全てのノードから成るベクトルを単に $\boldsymbol{x}$ と書くことにする．$\boldsymbol{x} = (\boldsymbol{y}^T, \boldsymbol{z}^T)^T$ である $(y_i = x_i,\ i = 1, \cdots, m)$, $(z_i = x_{i-m},\ i = m+1, \cdots, n)$．なお，^{T} は転置を表す．

グラフィカルモデルはノードとリンクからなるモデルである．リンクの集合を $\mathcal{L}$ とする．上の例であれば，$\mathcal{L} = \{(1,2),\ (2,3),\ (1,4),\ (2,5),\ (3,6),\ (4,5),\ (5,6),\ (3,7),\ (6,8)\}$ である．リンクに関しては添字の組のみで示すことにする．あるグラフィカルモデルの表現が与えられたとき，その確率変数に対する同時確率密度関数は

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

とかける．ここで Z は規格化定数であり $\psi_{ij}(x_i, x_j) > 0$ が常に成り立つとする．グラフィカルモデルには有向グラフ，無向グラフの区別があるが，Hammersley-Clifford 定理によって，全てのグラフィカルモデルはこのような無向グラフに変換できる．

各ノードの変数 x_i が 2 値の場合，$\psi_{ij}(x_i, x_j)$ は 2×2 で全ての成分が正である行列の形を取る．グラフィカルモデルの流儀によれば，確率密度関数の規格化に関しては別途扱うので，この時点では $\psi_{ij}(x_i, x_j)$ に関する規格化は考える必要がなく，定数倍についての差は同値として良い．ここで，観測されたノードの条件付きでの観測されていないノードの分布，すなわち $q(\boldsymbol{y}|\boldsymbol{z})$ を考える．

$$\begin{aligned} q(\boldsymbol{y}) &\stackrel{\text{def}}{=} q(\boldsymbol{y}|\boldsymbol{z}) \\ &= \frac{1}{Z} \prod_{(i,j) \in \mathcal{L}: i \leq m, j > m} \psi_{ij}(y_i, z_{j-m}) \prod_{(i,j) \in \mathcal{L}: i,j \leq m} \psi_{ij}(y_i, y_j) \\ &= \frac{1}{Z} \prod_{i=1}^{m} \varpi_i(y_i) \prod_{(i,j) \in \mathcal{L}: i,j \leq m} \psi_{ij}(y_i, y_j) \end{aligned} \tag{1}$$

ここで，

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

である．

2.2 確率伝搬法

今，あるノード y_i の確率分布が知りたいとする．これは，$q(\boldsymbol{y})$ を周辺化し y_i の確率分布を求めることに他ならない．なお $y_i \in \{-1, +1\}$ であるので y_i の期待値を計算することと確率分布を計算することは同値である．この周辺分布のことを “ビリーフ” と呼び，推定されたビリーフを $b_i(y_i)$ と書くことにする．真の分布の周辺化は $q_i(y_i)$ と書き，特に真のビリーフと呼ぶことにする．$q_i(y_i)$ は次の式で計算できる．

$$q_i(y_i) = \sum_{y_{i'}: i' \neq i} q(\boldsymbol{y}).$$

一般にはこの計算はノードの数に対して指数個の項の和が必要であり，NP-hard である．これに対し， J. Pearl は確率伝搬法と呼ばれる手法を提案し，グラフが “木” の構造をしている場合には有限回の操作で厳密な推定が可能であることを示した [9].

確率伝搬法では，ノード i から j への “メッセージ” と呼ばれるものが重要になる．i から j へのメッセージを $m_{ij}(y_j)$ と書く．このメッセージは ノード i から j への影響を y_j の確率分布の形で送るものである．無向グラフの場合にはリンクに向きが無いので，一つのリンク (i,j) に対して $m_{ij}(y_j)$ と $m_{ji}(y_i)$ を考える必要がある．確率伝搬法はこの $m_{ij}(y_j)$ の更新の方法である．以下に計算方法を示す．これからの議論のため，ノード i と継がっているノードの集合を $\mathcal{N}(i)$ と書くことにする．

確率伝搬法

1. $t = 0$ において，全てのメッセージに対し，$m^t_{ij}(y_j) = 1/2,\ (i,j) \in \mathcal{L}$ と初期値を与える．

2. $t = 1, 2, 3, \cdots$ に対して，以下の手続きを行ないメッセージとビリーフを更新する．

   (a) メッセージを次のように更新する．

   $$\begin{aligned} &m^{t+1}_{ij}(y_j) \\ &= \frac{1}{Z} \sum_{y_i} \varpi_i(y_i) \psi_{ij}(y_i, y_j) \prod_{k \in \mathcal{N}(i) \setminus j} m^t_{ki}(y_i) \end{aligned} \tag{2}$$

   Z は $\sum_{y_j} m^{t+1}_{ij}(y_j) = 1$ と規格化するための定数である．また，$\mathcal{N}(i) \setminus j$ は $\mathcal{N}(i)$ から j を除いたものである．

(b) ビリーフを次のように更新する．

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

3. 全ての $b_i(y_i)$ が収束したならば終了する．そうでなければ t を増やし，以上の手続きを繰り返す．

この確率伝搬法は，グラフの構造がループを持つ場合には必ずしもうまくいかないことが分っている．すなわち，必ずしも収束せず，また，収束したとしてもその解は真の周辺化 $q_i(y_i)$ とは等しくない．現実的にはループのあるようなグラフ構造から，幾つかのノードを合成することでループの無い構造を作り出し，この問題を避けることができる [6] が，規模の大きなグラフに対してはこの手法も必ずしもうまくいかない．確率伝搬法は規模の大きいグラフィカルモデルにおける周辺分布の近似手法として，1 つの有効な手法である．

3 情報幾何の準備

3.1 確率分布の空間，指数型分布族

本節では，本稿で必要となる情報幾何 [2] について述べる．まず，$\boldsymbol{y}$ の確率分布の空間を考える．今，各ノードは 2 値分布であるので，$\boldsymbol{y}$ の取り得る状態は 2^m 個である．したがって全ての $\boldsymbol{y}$ の分布は離散分布となる．この $\boldsymbol{y}$ の確率分布全体の集合を S と呼ぶことにする．

$$S = \left\{ p(\boldsymbol{y}) \middle| p(\boldsymbol{y}) > 0, \sum_{\boldsymbol{y}} p(\boldsymbol{y}) = 1 \right\}.$$

指数型分布族を以下のように定義する．

Definition 1. 密度関数が次のように定義される分布を指数型分布族と呼ぶ．

$$p(\boldsymbol{y};\boldsymbol{\theta}) = \exp\Big(k(\boldsymbol{y}) + \sum_{i=1}^{N}\theta^i F_i(\boldsymbol{y}) - \varphi(\boldsymbol{\theta})\Big)$$

ここで，$\boldsymbol{\theta} = (\theta^1,\cdots,\theta^N)^T$ は自然母数と呼ばれ，多様体 S の座標を与える．また，$F_i(\boldsymbol{y})$, $i = 1,\cdots,N$ は十分統計量である．

全ての離散分布は指数型分布族であるので [2]，S は指数分布族である．また，S の自由度は $2^m - 1$ である．

3.2 e–平坦，m–平坦な多様体と m–射影

次に，S の e–平坦，m–平坦な部分多様体を定義する．

e–平坦な部分多様体: 部分多様体 $M\in S$ は，任意の $q(\boldsymbol{y}), p(\boldsymbol{y}) \in M$ に対して次の式で定義される全ての $r(\boldsymbol{y};s)$ が M に属するとき，e–平坦であると呼ばれる．

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

ここで $c(s)$ は規格化定数である．定義より明らかに $r(\boldsymbol{y};s)$ は $p(\boldsymbol{y})$ と $q(\boldsymbol{y})$ とを結ぶ指数分布族を与える．特にある e–平坦な部分空間が 1 次元の自由度を持つとき，e–測地線と呼ぶ．上記の $r(\boldsymbol{y};s)$ は $p(\boldsymbol{y})$ と $q(\boldsymbol{y})$ とを結ぶ e–測地線である．

m–平坦な部分多様体: 部分多様体 $M\in S$ は，任意の $q(\boldsymbol{y}), p(\boldsymbol{y}) \in M$ に対して次の式で定義される全ての $r(\boldsymbol{y};s)$ が M に属するとき，m–平坦であると呼ばれる．

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

ある m–平坦な部分空間が 1 次元の自由度を持つとき，m–測地線と呼ぶ．したがって，上記の $r(\boldsymbol{y};s)$ は $p(\boldsymbol{y})$ と $q(\boldsymbol{y})$ の混合分布であり，m–測地線である．

次に m–射影と定義する [2].

Definition 2. M を S の e–平坦な部分多様体とする．今，S のある分布 $q(\boldsymbol{y})\in S$ に対して M 上で，次の *Kullback-Leibler* 情報量を最小にする点を考える．すなわち，

$$\Pi_M \circ q(\boldsymbol{y}) = \mathop{\mathrm{argmin}}_{p(\boldsymbol{y})\in M} D[q(\boldsymbol{y}); p(\boldsymbol{y})]. \tag{4}$$

この点を $q(y)$ から M への m–射影と呼ぶ．

なお，KL-情報量は次の式で定義される．

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

KL–情報量は $D[q(\boldsymbol{y}); p(\boldsymbol{y})]\geq 0$, を満し，$D[q(\boldsymbol{y}); p(\boldsymbol{y})] = 0$ が成り立つのは $q(\boldsymbol{y}) = p(\boldsymbol{y})$ が全ての $\boldsymbol{y}$ に対して成り立つときのみである．

なお，次の定理が示されている [2].

Theorem 1. m–射影は *1* 点に定まる．

3.3 ビリーフの表現

ここでビリーフのみから成る確率分布を考える．今，$y_i \in \{-1,+1\}$ であるので，$b_i(y_i)$ は 2 項分布であり，

したがって指数型分布族である．この指数分布族を次のように定める

$$\begin{aligned} b_i(y_i) \stackrel{\text{def}}{=} b_i(y_i;\theta^i) &= \varpi_i(y_i)\exp\bigl(\theta^i y_i - \phi_i(\theta^i)\bigr) \\ &= \exp(k_i(y_i) + \theta^i y_i - \phi_i(\theta^i)). \end{aligned} \tag{5}$$

ここで，$k_i(y_i) = \ln \varpi_i(y_i)$ とし，θ^i は自然母数であり，$\phi_i(\theta^i)$ は規格化定数である．

各 y_i のビリーフの積で表わされる分布を考える．

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

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

$$\boldsymbol{\theta} = (\theta^1, \cdots, \theta^m)^T, \boldsymbol{\theta} \in \mathcal{R}^m$$

$p_0(\boldsymbol{y};\boldsymbol{\theta})$ から成る多様体を M_0 とする．

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

$\boldsymbol{\theta}$ は M_0 の一つの座標系を与える．

別の座標系として，期待値パラメータ $\boldsymbol{\eta}_0 = (\eta_{01}, \cdots, \eta_{0m})^T$ を次のように定義する．

$$\boldsymbol{\eta}_0 = E_{p_0(\boldsymbol{y};\boldsymbol{\theta})}[\boldsymbol{y}] = \sum_{\boldsymbol{y}} p_0(\boldsymbol{y};\boldsymbol{\theta})\boldsymbol{y} = \partial_{\boldsymbol{\theta}}\varphi_0(\boldsymbol{\theta})$$

$E_p[\cdot]$ は p による平均を意味する．$\boldsymbol{\theta}$ と $\boldsymbol{\eta}_0$ は 1 対 1 の関係であり，特に $\boldsymbol{\eta}_0$ を $\boldsymbol{\theta}$ の関数として考えるときには $\boldsymbol{\eta}_0(\boldsymbol{\theta})$ と書くことにする．

Proposition 1. **M_0 は e–平坦な多様体である．**

Proof. $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathcal{R}^N$ と置き，$p_0(\boldsymbol{x};\boldsymbol{\theta}), p_0(\boldsymbol{x};\boldsymbol{\theta}') \in M_0$ とする．全ての $\boldsymbol{\theta}, \boldsymbol{\theta}'$ に対し，

$$\begin{aligned} \ln r(\boldsymbol{y};s) &= (1-s)\ln p_0(\boldsymbol{y};\boldsymbol{\theta}) + s\ln p_0(\boldsymbol{y};\boldsymbol{\theta}') + c(s) \\ &= k_0(\boldsymbol{y}) + \boldsymbol{u}(s)\cdot\boldsymbol{y} - \varphi_0(\boldsymbol{u}(s)). \end{aligned}$$

となり M_0 に属することがわかる． なお，ここで $\boldsymbol{u}(s) \stackrel{\text{def}}{=} (1-s)\boldsymbol{\theta} + s\boldsymbol{\theta}'$, $c(s) = (1-s)\varphi_0(\boldsymbol{\theta}) + s\varphi_0(\boldsymbol{\theta}') - \varphi(\boldsymbol{u}(s))$ とおいた． □

確率伝搬法の目的は，この M_0 の中から，真の分布の周辺化となる一点を求めることである．

3.4 真の分布とビリーフ

(1) 式にある真の分布 $q(\boldsymbol{y})$ は次のように書きなおせる．

$$\begin{aligned} q(\boldsymbol{y}) &= \frac{1}{Z}\prod_{i=1}^{m}\varpi_i(y_i)\prod_{(i,j)\in\mathcal{L}}\psi_{ij}(y_i,y_j) \\ &= C\exp\Bigl(k_0(\boldsymbol{y}) + \sum_{(i,j)\in\mathcal{L}}\ln\psi_{ij}(y_i,y_j)\Bigr) \qquad (6) \\ &= C\exp\Bigl(k_0(\boldsymbol{y}) + \sum_{r=1}^{L} c_r(\boldsymbol{y})\Bigr) \end{aligned}$$

ここで，$c_r(\boldsymbol{y})$ は $\ln\psi_{ij}(y_i,y_j)$ に対応し，r は (i,j) について添字を付け直したものである．L は $\mathcal{L}$ の成分の数，すなわちリンクの数である．

グラフィカルモデルの推論の問題は (6) によって定義される $q(\boldsymbol{y})$ に対して，ある M_0 の一点を求めることとなる．

Proposition 2. **真のビリーフは，$q(\boldsymbol{y})$ から M_0 への m–射影と同値である．**

Proof. 定義より m–射影は $q(\boldsymbol{y})$ から $p_0(\boldsymbol{y};\boldsymbol{\theta})$ への Kullback-Leibler 情報量 $D[q(\boldsymbol{y});p_0(\boldsymbol{y};\boldsymbol{\theta})]$ を最小にするパラメータを求めることである．M_0 は e–平坦なので，最小にする点は一点のみである．KL–情報量の $\boldsymbol{\theta}$ による微分は

$$\begin{aligned} \partial_{\boldsymbol{\theta}} D[q(\boldsymbol{y});p_0(\boldsymbol{y};\boldsymbol{\theta})] &= \partial_{\boldsymbol{\theta}}\varphi_0(\boldsymbol{\theta}) - \sum_{\boldsymbol{x}} q(\boldsymbol{y})\boldsymbol{y} \\ &= \boldsymbol{\eta}_0(\boldsymbol{\theta}) - E_{q(\boldsymbol{y})}[\boldsymbol{y}]. \end{aligned}$$

m–射影によって求まる点ではこの微分が 0 となる．したがって，射影された点の θ–座標を $\boldsymbol{\theta}^*$ とすれば，

$$\boldsymbol{\eta}_0(\boldsymbol{\theta}^*) = E_{q(\boldsymbol{y})}[\boldsymbol{y}] = \sum_{\boldsymbol{y}} q(\boldsymbol{y})\boldsymbol{y}$$

である．これは，$b_i(y_i;\theta^i)$ が正しく $q_i(y_i)$ を表現していることを示している．したがって $q(\boldsymbol{y})$ の M_0 への m–射影は $\boldsymbol{y}$ の期待値を変えず，$q(\boldsymbol{y})$ の各ノードの分布への周辺化と同値であることが分る (Fig.2)．すなわち

$$p_0(\boldsymbol{y};\boldsymbol{\theta}^*) = \mathit{\Pi}_{M_0} \circ q(\boldsymbol{y})$$

である． □

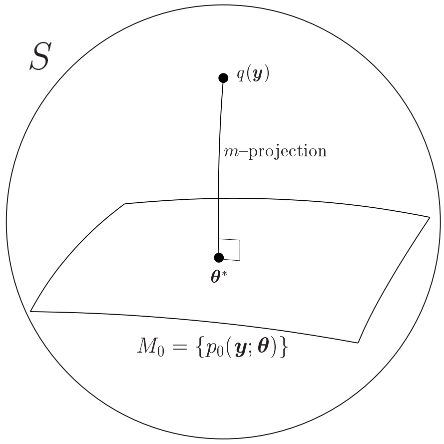


図 2: m–射影と周辺化

また，ここで特に多様体 M_0 への m–射影によってパラメータを求めるオペレータを π_{M_0} とおく，

$$\pi_{M_0} \circ q(\boldsymbol{y}) \stackrel{\text{def}}{=} \operatorname*{argmin}_{\boldsymbol{\theta}} D[q(\boldsymbol{y});p_0(\boldsymbol{y};\boldsymbol{\theta})]. \tag{7}$$

定義より

$$\boldsymbol{\theta}^* = \pi_{M_0} \circ q(\boldsymbol{y})$$

である．

4　確率伝搬法の情報幾何

4.1　メッセージの情報幾何的表現

本節では，確率伝搬法におけるメッセージを情報幾何的に書き直す．(3) 式がメッセージとビリーフの関係を定義している．(5) 式の結果と合わせて，

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

とかける．$m_{ki}(y_i)$ はノード k から i へ確率分布を伝搬させており，k を固定すると $m_{ki}(y_i)$ は 2 個の成分を持つ．しかし，確率分布として和が 1 である条件を考え合わせると，自由度は 1 しかない．ここで μ_k^i を次のように定義すると，

$$\frac{m_{ki}(y_i)}{\sum_{y_i} m_{ki}(y_i)} = \frac{1}{2}(\tanh(\mu_k^i y_i)+1)$$

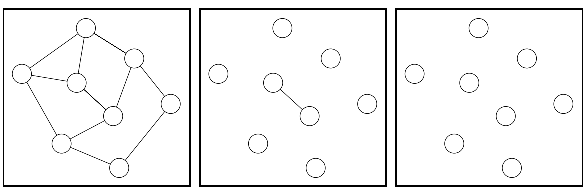
μ_k^i と $m_{ki}(y_i)$ とが 1 対 1 に対応することがわかる．したがって，確率伝搬法では，メッセージを用い，パラメータを

$$\theta^i = \sum_{k\in\mathcal{N}(i)} \mu_k^i \tag{8}$$

と定義していることと等しい．この結果を用いて，ビリーフは次のようにかける．

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

4.2　リンクと確率分布



A　　　B　　　C

図 3: 3 つの重要なグラフ, A: 全てのリンクがあるグラフ，真の分布．B: 確率伝搬法で重要となる一つだけリンクを持つグラフ，C: ビリーフのみから成るリンクのないグラフ

同じノードの組に対し，確率伝搬法で重要となる 3 つのグラフを示す (図 3)．図 3.A は真の分布，すなわち $q(\boldsymbol{y})$ を表わす．グラフィカルモデルにおいて，特に確率伝搬法を考えるときには各ノードの確率分布を個別に考えることが多いが，我々はいま S の空間での情報幾何を考えており，その意味で 1 つ 1 つのノードを別々に考えるのではなく，全てのノードの同時分布を考える方が良い．図 3.C は各ノードのビリーフのみを考え，その間の相互作用はないとしたときのモデルである．これは M_0 の点となる．

では，(2) 式のように，1 つのリンクのみを考慮にいれたものはどうなるだろうか．リンクの両端のノード以外は ビリーフ で置き換えると，グラフは図 3.B のようになる．

仮に $\psi_{ij}(y_i,y_j)$ のリンクを考慮したとしよう．このとき，y_i, y_j 以外のノードは $b_i(y_i;\theta^i)$ を用いるものとする．$\psi_{ij}(y_i,y_j)$ は 2 つのメッセージの代わりに用いられる．すなわち，$m_{ij}(y_j)$ と $m_{ji}(y_i)$ である．これに対応する $\{\mu_i^j\}$ は μ_i^j と μ_j^i である ($i<j$ とする)．そこで，次のベクトル $\boldsymbol{\xi}_r$ を次のように定義する．r は (6) 式と同様に (i,j) に番号を振り直したものである．

$$\boldsymbol{\xi}_r = (0,\cdots,0,\underset{\stackrel{\wedge}{i}}{\mu_j^i},0,\cdots,0,\underset{\stackrel{\wedge}{j}}{\mu_i^j},0,\cdots,0)^T \tag{9}$$

(8) 式の結果から，次式が成り立つ．

$$\boldsymbol{\theta} = \sum_{r\in\mathcal{L}} \boldsymbol{\xi}_r.$$

これを用いると，図 3.B のモデルは，

$$\begin{aligned} p_r(\boldsymbol{y};\boldsymbol{\zeta}_r) &\stackrel{\mathrm{def}}{=} p_r(\boldsymbol{y};\boldsymbol{\theta}-\boldsymbol{\xi}_r) \\ &= \frac{1}{Z}\psi_{ij}(y_i,y_j)p_0(\boldsymbol{y};\boldsymbol{\theta}-\boldsymbol{\xi}_r) \\ &= \exp(k_0(\boldsymbol{y})+c_r(\boldsymbol{y})+\boldsymbol{\zeta}_r\cdot\boldsymbol{y}-\varphi_r(\boldsymbol{\zeta}_r)) \end{aligned} \tag{10}$$

とかける．ここで $\boldsymbol{\zeta}_r \stackrel{\mathrm{def}}{=} \boldsymbol{\theta}-\boldsymbol{\xi}_r$ である．

このように 1 つだけのリンクを入れたものをリンクの数だけ考える必要がある．それぞれの確率分布の多様体を M_r と定義する．

$$M_r = \Bigl\{p_r(\boldsymbol{y};\boldsymbol{\zeta}_r) = \exp(k_0(\boldsymbol{y})+c_r(\boldsymbol{y})+\boldsymbol{\zeta}_r\cdot\boldsymbol{y}-\varphi_r(\boldsymbol{\zeta}_r)) \Bigm| \boldsymbol{\zeta}_r\in\mathcal{R}^N\Bigr\},\quad r=1,\cdots,L. \tag{11}$$

$\boldsymbol{\zeta}_r$ は M_r の自然母数である．M_r もまた定義より e–平坦な多様体である．しかし，一般に $M_r \neq M_0$ であり，$r \neq r'$ に対しては $c_r(\boldsymbol{y})\neq c_{r'}(\boldsymbol{y})$ であることから

$M_r{\neq}M_{r'}$, である．$p_r(\boldsymbol{y};\boldsymbol{\zeta}_r)$ の期待値パラメータは以下のように定義される．

$$\boldsymbol{\eta}_r(\boldsymbol{\zeta}_r)\stackrel{\rm def}{=}\sum_{\boldsymbol{y}} p_r(\boldsymbol{y};\boldsymbol{\zeta}_r)\boldsymbol{y}=\partial_{\boldsymbol{\zeta}_r}\varphi_r(\boldsymbol{\zeta}_r),\quad r=1,\cdots,L. \tag{12}$$

4.3 確率伝搬法の情報幾何的表現

本節では確率伝搬法を情報幾何的に書き直す．(2) 式の両辺に $\varpi_j(y_j)\prod_{k\in\mathcal{N}(j)\setminus i} m^t_{kj}(y_j)$ を乗じる．

$$\begin{aligned}&\frac{1}{Z'}m^{t+1}_{ij}(y_j)\varpi_j(y_j)\prod_{k\in\mathcal{N}(j)\setminus i} m^t_{kj}(y_j)\\ =&\frac{1}{Z}\sum_{y_i}\varpi_i(y_i)\varpi_j(y_j)\psi_{ij}(y_i,y_j)\prod_{k\in\mathcal{N}(j)\setminus i}m^t_{kj}(y_j)\prod_{k\in\mathcal{N}(i)\setminus j}m^t_{ki}(y_i)\end{aligned}$$

$m^t_{ij}(y_j)$ が $\boldsymbol{\xi}^t_r$ (r は (i,j) に対応している) と対応していることを用いると，左辺は,

$$\sum_{y_k:k\neq j} p_0(\boldsymbol{y};\boldsymbol{\theta}^t-\boldsymbol{\xi}^t_r+\boldsymbol{\xi}^{t+1}_r)=\sum_{y_k:k\neq j} p_0(\boldsymbol{y};\boldsymbol{\zeta}^t_r+\boldsymbol{\xi}^{t+1}_r)$$

とかける．右辺は $p_r(\boldsymbol{y};\boldsymbol{\zeta}_r)$ を y_j に関して周辺化したものに等しいことが分る．すなわち，

$$\sum_{y_k:k\neq j} p_0(\boldsymbol{y};\boldsymbol{\zeta}^t_r+\boldsymbol{\xi}^{t+1}_r)=\sum_{y_k:k\neq j} p_r(\boldsymbol{y};\boldsymbol{\zeta}_r).$$

が成り立つ．この結果を用いると，(2) 式が次のように書けることが分る．

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

これを用いて確率伝搬法の情報幾何的表現を次のように得る．

確率伝搬法

1. $t=0$ において，$\boldsymbol{\theta}=\mathbf{o}$, $\boldsymbol{\xi}_r=\mathbf{o}$, $r=1,\cdots,L$ とする．

2. $t=1,2,3,\cdots$ に対して，以下の手続きを行ないメッセージとビリーフ を更新する．

   (a) $\boldsymbol{\xi}^{t+1}_r$, $r=1,\cdots,L$ に対しては次のように更新する．

   $$\boldsymbol{\xi}^{t+1}_r=\pi_{M_0}\circ p_r(\boldsymbol{y};\boldsymbol{\zeta}^t_r)-\boldsymbol{\zeta}^t_r. \tag{13}$$

   (b) $\boldsymbol{\theta}$, $\boldsymbol{\zeta}_r$, $r=1,\cdots,L$ を次のように更新する．

   $$\boldsymbol{\theta}^{t+1}=\sum_{r=1}^L\boldsymbol{\xi}^{t+1}_r,\quad \boldsymbol{\zeta}^{t+1}_r=\boldsymbol{\theta}^{t+1}-\boldsymbol{\xi}^{t+1}_r. \tag{14}$$

3. $\boldsymbol{\xi}^{t+1}_r$ が収束したならば終了する．そうでなければ t を 1 つ増やし，以上の手続きを繰り返す．最終的なビリーフは収束値を ˆ で表すことにすると，$p_0(\boldsymbol{y};\hat{\boldsymbol{\theta}})$ である．

---

5 確率伝搬法の停留点の満たす性質

確率伝搬法の停留点の満たす条件は次の 2 つの式で示される．

1) $\Pi_{M_0}\circ p_r(\boldsymbol{y};\hat{\boldsymbol{\zeta}}_r)=p_0(\boldsymbol{y};\hat{\boldsymbol{\theta}}), r=1,\cdots,L.$
 期待値パラメータを用いて書き直すと

$$\boldsymbol{\eta}_0(\hat{\boldsymbol{\theta}})=\boldsymbol{\eta}_1(\hat{\boldsymbol{\zeta}}_1)=\cdots=\boldsymbol{\eta}_L(\hat{\boldsymbol{\zeta}}_L). \tag{15}$$

2) $$\hat{\boldsymbol{\theta}}=\sum_{r=1}^L\hat{\boldsymbol{\xi}}_r=\frac{1}{L-1}\sum_{r=1}^L\hat{\boldsymbol{\zeta}}_r. \tag{16}$$

ここで次の 2 つの多様体を考える．まず，$M(\boldsymbol{\theta})$ は次のように定義される m–平坦な多様体である．

$$M(\boldsymbol{\theta})=\Big\{p(\boldsymbol{y})\mid p(\boldsymbol{y})\in S,\ \sum_{\boldsymbol{y}}p(\boldsymbol{y})\boldsymbol{y}=\boldsymbol{\eta}_0(\boldsymbol{\theta})\Big\}.$$

全ての $p(\boldsymbol{y}){\in}M(\boldsymbol{\theta})$ に対して $\boldsymbol{y}$ の期待値は同じであり，したがって $p(\boldsymbol{y}){\in}M(\boldsymbol{\theta})$, の M_0 への m–射影は $p_0(\boldsymbol{y};\boldsymbol{\theta})$ となる．

また，$M(\boldsymbol{\theta})$ に含まれる各多様体 M_r $r=1,\cdots,L$ の点を $\boldsymbol{\zeta}_r(\boldsymbol{\theta})$ と書くことにする．

$$\boldsymbol{\zeta}_r(\boldsymbol{\theta})\stackrel{\rm def}{=}\pi_{M_r}{\circ}p_0(\boldsymbol{x};\boldsymbol{\theta}),\quad r=1,\cdots,L.$$

M_r が e–平坦であることと，$\boldsymbol{\eta}_r(\boldsymbol{\zeta}_r)=\partial_{\boldsymbol{\zeta}_r}\varphi_r(\boldsymbol{\zeta}_r)$ より，$p_0(\boldsymbol{y};\boldsymbol{\theta})$ の M_r 上への m–射影が $\boldsymbol{\zeta}_r(\boldsymbol{\theta})$ である．

次に，e–平坦な多様体 $E(\boldsymbol{\theta})$ を定義する．これは $p_0(\boldsymbol{x};\boldsymbol{\theta})$ と $p_r(\boldsymbol{x};\boldsymbol{\zeta}_r(\boldsymbol{\theta}))$, $r=1,\cdots,L$ とを log-linear に結ぶ多様体であり，定義は次の通りである．

$$\begin{aligned}E(\boldsymbol{\theta})=\Big\{p(\boldsymbol{y})\mid p(\boldsymbol{y})=Cp_0(\boldsymbol{y};\boldsymbol{\theta})^{t_0}\prod_{r=1}^L p_r(\boldsymbol{y};\boldsymbol{\zeta}_r(\boldsymbol{\theta}))^{t_r},\\ t_r{\in}\mathcal{R},\ \sum_{r=0}^L t_r=1\Big\}\ C:\text{normalization factor},\end{aligned}$$

明らかに $p_0(\boldsymbol{x};\boldsymbol{\theta})$, $p_r(\boldsymbol{x};\boldsymbol{\zeta}_r(\boldsymbol{\theta}))$, $r=1,\cdots,L$ は全て $E(\boldsymbol{\theta})$ に含まれる．さらに，アルゴリズムの収束点では，(16) が満たされるので，$t_0=-(L-1)$, $t_1=,\cdots,=t_L=1$ とおくと，

$$C\frac{\prod_{r=1}^L p_r(\boldsymbol{y};\hat{\boldsymbol{\zeta}}_r)}{p_0(\boldsymbol{x};\hat{\boldsymbol{\theta}})}=C\exp\Big(k_0(\boldsymbol{y})+\sum_{r=1}^L c_r(\boldsymbol{y})\Big)=q(\boldsymbol{y}).$$

したがって，条件 2) によって，$q(\boldsymbol{y})$ もまた e–平坦な多様体 $E(\hat{\boldsymbol{\theta}})$ に含まれていることが分る．この結果，次の定理が導かれる．

Theorem 2. **確率伝搬法の収束点では** $p_0(\boldsymbol{y};\hat{\boldsymbol{\theta}})$, $p_r(\boldsymbol{y};\hat{\zeta}_r)$, $(r=1,\cdots,L)$ **が** m–**平坦な多様体** $M(\hat{\boldsymbol{\theta}})$ **に含まれ，**e–**平坦な多様体** $E(\hat{\theta})$ **は** $p_0(\boldsymbol{x};\boldsymbol{\theta}^*)$, $p_r(\boldsymbol{y};\hat{\zeta}_r)$, $(r=1,\cdots,L)$ **と** $q(\boldsymbol{x})$ **を含む．**

この定理によって，確率伝搬法の収束点の幾何的な性質が明らかになる．もし，$M(\hat{\boldsymbol{\theta}})$ が $q(\boldsymbol{y})$ を含めば $p(\boldsymbol{y};\hat{\boldsymbol{\theta}})$ が $q(\boldsymbol{y})$ の真の周辺化となる．しかし，一般には $M(\hat{\boldsymbol{\theta}})$ は必ずしも $q(\boldsymbol{y})$ を含まない．$M(\hat{\boldsymbol{\theta}})$, の代わりに e–平坦な多様体 $E(\hat{\boldsymbol{\theta}})$ が $q(\boldsymbol{y})$ を含むことで，確率伝搬法は真の周辺化の近似解を与えている．このような構造は平均場近似の構造と似ている [1, 5, 10]．一般には $M(\hat{\boldsymbol{\theta}})$ と $E(\hat{\boldsymbol{\theta}})$, の間には差がり，これが確率伝搬法の誤差となる．したがってこの差を見積もれば，確率伝搬法の解を改善できる可能性もある．

また，グラフが木の構造をしている場合には次の関係が成り立つ．

Theorem 3. **グラフの構造が木の構造であるときには，真の分布** $q(\boldsymbol{y})$ **は次のように構成できる．**

$$q(\boldsymbol{y})=\frac{\prod_{(i,j)\in\mathcal{L}}q_{ij}(y_i,y_j)}{\prod_{i'=1}^{m}q_{i'}(y_{i'})^{n_{i'}-1}} \tag{17}$$

ここで n_i **はノード** i **につながっているノードの数であり，**$q_{ij}(y_i,y_j)$ **は** (y_i,y_j) **について** $q(\boldsymbol{y})$ **を周辺化したものである．**

この定理は情報幾何的には，どのように理解されるだろうか．$\prod_{i=1}^m q_i(y_i)$ は M_0 含まれる．そのときのパラメータは $\boldsymbol{\theta}^*$ である．一方 $q_{ij}(y_i,y_j)$ に $\prod_{k=1}^m q_k(y_k)/(q_i(y_i)q_j(y_j))$ を乗じた分布は M_r に含まれる．このパラメータを $\boldsymbol{\zeta}'_r$ とする．すなわち，

$$p_0(\boldsymbol{y};\boldsymbol{\theta}^*)=\prod_{i=1}^m q_i(y_i)$$
$$p_r(\boldsymbol{y};\boldsymbol{\zeta}'_r)=q_{ij}(y_i,y_j)\frac{\prod_{i'=1}^m q_{i'}(y_{i'})}{q_i(y_i)q_j(y_j)},\quad r=1,\cdots,L$$

明らかに $p_r(\boldsymbol{y};\boldsymbol{\zeta}'_r)$ の周辺化は $p_0(\boldsymbol{y};\boldsymbol{\theta}^*)$ である．したがって，$M(\boldsymbol{\theta}^*)$ に $p_0(\boldsymbol{y};\boldsymbol{\theta}^*)$, $p_r(\boldsymbol{y};\boldsymbol{\zeta}'_r)$, $r=1,\cdots,L$, そして $q(\boldsymbol{y})$ が含まれる．いま，木構造のグラフィカルモデルでは，Theorem 3 からさらに次の結果がえられる．

Proposition 3. **木構造のグラフでは，**$p_0(\boldsymbol{y};\boldsymbol{\theta}^*)$, $p_r(\boldsymbol{y};\boldsymbol{\zeta}'_r)$, $r=1,\cdots,L$, **そして** $q(\boldsymbol{y})$ **が** $M(\boldsymbol{\theta}^*)$ **そして** $E(\boldsymbol{\theta}^*)$ **に含まれる．**

Proof. 上の定義より $M(\boldsymbol{\theta}^*)$ に含まれることは明らかである．

$$\begin{aligned}\prod_{(i,j)\in\mathcal{L}}p_r(\boldsymbol{y}_r;\boldsymbol{\zeta}'_r) &= \prod_{(i,j)\in\mathcal{L}}\frac{q_{ij}(y_i,y_j)\prod_{k=1}^m q_k(y_k)}{q_i(y_i)q_j(y_j)}\\ &= \prod_{(i,j)\in\mathcal{L}}q_{ij}(y_i,y_j)\prod_{k=1}^m q_k(y_k)^{L-n_k}\end{aligned}$$
$$\prod_{(i,j)\in\mathcal{L}}q_{ij}(y_i,y_j)=\frac{\prod_{(i,j)\in\mathcal{L}}p_r(\boldsymbol{y}_r;\boldsymbol{\zeta}'_r)}{\prod_{k=1}^m q_k(y_k)^{L-n_k}}$$

(17) 式の結果に代入すると，

$$\frac{\prod_{(i,j)\in\mathcal{L}}q_{ij}(y_i,y_j)}{\prod_{i'=1}^m q_{i'}(y_{i'})^{n_{i'}-1}}=\frac{\prod_r p_r(\boldsymbol{y}_r;\boldsymbol{\zeta}'_r)}{p_0(\boldsymbol{y};\boldsymbol{\theta}^*)^{L-1}}=q(\boldsymbol{y})$$

したがって $E(\boldsymbol{\theta}^*)$ に含まれる． □

この結果より木の構造のグラフでは確率伝搬法が正しい答えをあたえることがわかる．

6 まとめ

本稿では，情報幾何によって確率伝搬法を表現した結果を示した．我々はこの枠組みの下で，確率伝搬法のさらなる解析，拡張を論じていくことができると考えている．本稿が確率伝搬法の理解，そしてさらなる研究のための助けになればと思う．

我々はターボ符号，低密度パリティ検査符号の繰り返し復号法に対する同様の枠組みを示し，繰り返し復号法の幾つかの性質を示した．たとえば，局所的なアルゴリズムの停留点の安定性，また，真の解からの誤差の解析を行なった．

これらの結果は，多少の拡張は必要だが本稿で示した一般の確率伝搬法に対しても示すことができる．今後拡張した結果についても機会を見て示していきたい．また，近年，確率伝搬法に関係する様々な手法が提案されている．これらは収束性，またアルゴリズムの結果得られる収束点の性質が確率伝搬法のそれとは異なるものの，やはり情報幾何的な枠組みによって表現することができる．今後，この枠組みから確率伝搬法の改善手法を提案できればと考えている．

参考文献

[1] Shun-ichi Amari, Shiro Ikeda, and Hidetoshi Shimokawa. Information geometry and mean field approximation: The α-projection approach. In Manfred Opper and David Saad, editors, *Advanced Mean Field Methods – Theory and Practice*, chapter 16, pages 241–257. The MIT Press, 2001.

[2] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*. AMS and Oxford University Press, 2000.

[3] Shiro Ikeda, Toshiyuki Tanaka, and Shun-ichi Amari. Information geometrical framework for analyzing belief propagation decoder. In Thomas G. Dietterich, Sue Becker, and Zoubin Ghahramani, editors, *Advances in Neural Information Processing Systems 14*. The MIT Press, Cambridge, MA, 2002.

[4] Yoshiyuki Kabashima and David Saad. Belief propagation vs. TAP for decoding corrupted messages. *Europhysics Letters*, 44(5):668–674, December 1998.

[5] Yoshiyuki Kabashima and David Saad. The TAP approach to intensive and extensive connectivity systems. In Manfred Opper and David Saad, editors, *Advanced Mean Field Methods – Theory and Practice*, chapter 6, pages 65–84. The MIT Press, 2001.

[6] Steffen L. Lauritzen and David 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.

[7] David J. C. MacKay. Good error-correcting codes based on very sparse matrices. *IEEE Transactions on Information Theory*, 45(2):399–431, March 1999.

[8] Robert J. McEliece, David J. C. MacKay, and Jung-Fu Cheng. Turbo decoding as an instance of Pearl's "belief propagation" algorithm. *IEEE Journal on Selected Areas in Communications*, 16(2):140–152, February 1998.

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

[10] Toshiyuki Tanaka. Information geometry of mean-field approximation. In Manfred Opper and David Saad, editors, *Advanced Mean Field Methods – Theory and Practice*, chapter 17, pages 259–273. The MIT Press, 2001.

[11] Toshiyuki Tanaka, Shiro Ikeda, and Shun-ichi Amari. Information-geometrical significance of sparsity in Gallager codes. In Thomas G. Dietterich, Sue Becker, and Zoubin Ghahramani, editors, *Advances in Neural Information Processing Systems 14*. The MIT Press, Cambridge, MA, April 2002.