確率伝搬法の情報幾何
– 符号理論，統計物理，人工知能の接点 –

池田 思朗, 統計数理研究所
田中 利幸, 東京都立大学
甘利 俊一, 理化学研究所

平成16年 6月 2日

1　はじめに

ここ数年確率伝搬法 (belief propagation) という手法が注目され，理論的な側面からの研究が盛んに行われている．確率伝搬法は 1980 年代後半に確率にもとづく推論の手法として J. Pearl によって提案されたものであり [16]，Pearl's belief propagation と呼ばれることもある．分類するならば，人工知能の手法として提案されたとするのが適切だろう [18]．この手法は局所的な情報のみを用いた計算を繰り返し行い，大域的な最適解を得ようというものであり，ある種の問題では非常に効率よく正しい解が得られることがわかっている．

近年注目されているのはこの「解ける」問題ではなく，この手法では「厳密解は得られない」と思われている問題に対して確率伝搬法が非常に良い振る舞いをする点である．理論的にはどの程度うまくいくのかわからないが，現実的には多くの場合でよい結果が得られるのだ．興味深いことに，同等の手法は符号理論，統計物理の分野でも古くから提案され，用いられている．したがって，確率伝搬法がなぜそのような問題に対してうまく働くのかを明らかにすることは，人工知能に限らず符号理論，統計物理の分野からも興味ある問題である．

当然，各分野において様々な面から関連する研究が行われてきた．しかし，満足できる理論的結果は現在までのところ得られていない．その中において，我々も情報幾何 [1] の立場からこの問題に取り組んでいる [6, 7, 8, 9]．我々の興味は各分野で得られている知見を統一的に扱える数理的枠組みを提案し，その上で各分野で既に得られている結果をまとめること，そしてこの不思議な手法の本質を明らかにすることである．

本稿では確率伝搬法がどのような手法なのか，各分野の手法を紹介し，我々が提案している情報幾何の枠組みと得られた結果，今後の展望について述べる．

2　確率伝搬法と関連する研究

2.1　確率伝搬法

確率伝搬法はベイジアンネット，あるいはグラフィカルモデルと呼ばれるグラフで表現される確率分布の推論に用いられる．図 1 に簡単なグラフイカルモデルを示す．グラフィカルモデルでは各確率変数を節によって表し，節の間を結ぶ辺はそれらの確率変数の間に直接の相互作用があることを表す．なお，各辺に向き (通常は矢印により示す) を与え確率変数間の関係を定義する有向グラフと，向きの無い無向グラフの 2 つの立場がある．本稿では無向グラフのみを扱うが，全ての有向グラフは無向グラフで表現できる [13] ことから一般性は失っていない．各確率変数はそれぞれの問題によって様々な意味が与えられる．人工知能の問題，例えば診断に関するグラフであれば，いくつかの変数は検査項目を示し，いくつかは病気の可能性を示すだろう．統計物理ならば磁性体のスピンの向き，誤り訂正符号の問題ならばビットの値が確率変数である．

グラフィカルモデルでは，$\boldsymbol{x} = (x_1, \cdots, x_N)^T$ の同時分布が

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

と定義される．なお，$\mathcal{L} = \{(i,j)\}$ は辺の集合，Z は和を 1 に規格化するための定数である．後の議論のためにより一般的な定義を以下に与えておく．

$$q(\boldsymbol{x}) = \frac{1}{Z}\prod_i \psi_i(x_i)\prod_{r\in\mathcal{L}}\psi_r(\boldsymbol{x}). \tag{2}$$

ψ_r はクリーク関数と呼ばれ，正の値をとる．特に図 1 のように 2 つの節の間の相互作用のみを考える場合はクリーク関数は辺で結ばれた確率変数間のみに定義され $\psi_r(\boldsymbol{x}) = \psi_{ij}(x_i, x_j)$ となり，(2) 式は (1) 式となる．

確率伝搬法はこの同時分布から各確率変数の周辺分布 $q(x_i)$ を効率良く求めるための手法である．前に示した問題において各確率変数の周辺分布が重要なことは明らかであろう．以下では簡単のため x_i が $\{-1,+1\}$ の 2 値をとり，周辺分布がベルヌーイ分布だとするが，多値をとる多項分布への拡張は簡単である．また連続値を取る場合でも，周辺分布が指数型分布族となる場合への拡張は可能である [9]．単純に考えると $q(x_i) = \sum_{x_j:j\neq i} q(\boldsymbol{x})$ を求めるには 2^{N-1} の要素の和が必要である．Z を求めることを考えれば 2^N の要素の和が必要となる．一方，確率伝搬法では辺の数に比例した回数の計算から周辺分布を求める．

2 つの節の間の相互作用のみがあるグラフィカルモデルに対する確率伝搬法を，[20] にしたがって定義しておく．確率伝搬法ではメッセージ $m_{ji}(x_i)$ を次の式で更新し，全てが収束した後，得られたメッセージ $m^*_{ji}(x_i)$ を用いてビリーフ $b_i(x_i)$ を求める．

$$\begin{aligned} m^{t+1}_{ji}(x_i) =& \frac{1}{Z}\sum_{x_j}\psi_j(x_j)\psi_{ij}(x_i,x_j)\prod_{k\in\mathcal{N}(j)\setminus i} m^t_{kj}(x_j) \\ b_i(x_i) =& \frac{1}{Z'}\psi_i(x_i)\sum_{k\in\mathcal{N}(i)} m^*_{ki}(x_i). \end{aligned} \tag{3}$$

$\mathcal{N}(j)$ は j 番目の節につながっている節の集合，Z，Z' はそれぞれ $m^{t+1}_{ji}(x_i)$ および $b_i(x_i)$ の x_i に関する和を 1 に規格化する．ビリーフ $b_i(x_i)$ は x_i の周辺分布を表す．

上に示した更新則では $\sum_{x_j}\psi_j(x_j)\psi_{ij}(x_i,x_j)$ において x_j に関する和をとっている．この計算を辺の数だけ行い，メッセージが更新される．大きなグラフに対して確率伝搬法が必要とする計算量は，周辺分布を直接計算するのに比べて格段に少ない．特に木のグラフ，すなわちループの無いグラフに対しては，メッセージの更新を辺の数だけグラフの端から端まで往復させて行えば，正しく周辺分布が求まることが示されている [16]．

Pearl の定式化により様々な手法が木のグラフに対する確率伝搬法であることがわかった．例えば音声認識で用いられる HMM(hidden Markov model) の forward-backward アルゴリズム，誤り訂正符号で用いられる BCJR (Bahl-Cocke-Jelinek-Raviv) アルゴリズム [2] も確率伝搬法として理解できる．

一方，ループのあるグラフに確率伝搬法を用いるとどうなるだろう．ループのある場合「端から端まで」確率を伝搬させることはできず，更新の手続きがぐるぐると回ってしまうことになる．実際にループのあるグラフに確率伝搬法を用いると，うまくいっても何回かの繰り返しが必要になり，場合によっては

収束しないこともある．また得られた結果が真の周辺分布であるという保証はない．それでもなお，確率伝搬法は現実的に有効な手段であり，実験的には多くの場合良い結果を与えることが報告されている．

2.2 問題の定式化

我々の目的は，この不思議な計算手法を情報幾何に基づき理解することである．準備のために (2) 式を以下のように再定義する．

$$\begin{aligned} q(\boldsymbol{x}) &= \frac{1}{Z}\prod_i \psi_i(x_i)\prod_{r\in\mathcal{L}}\psi_r(\boldsymbol{x}) = C\exp\Big(c_0(\boldsymbol{x}) + \sum_{r=1}^{L} c_r(\boldsymbol{x})\Big) \\ c_0(\boldsymbol{x}) &= \sum_i \ln\psi_i(x_i) + C' \quad c_r(\boldsymbol{x}) = \ln\psi_r(\boldsymbol{x}). \end{aligned} \tag{4}$$

$c_0(\boldsymbol{x})$ は $x_i \in \{-1,+1\}$ のとき，定数の差を除いて $\boldsymbol{\alpha}\cdot\boldsymbol{x}$, $\boldsymbol{\alpha}\in\Re^N$ とかける．ただし $\cdot$ は内積であり，$\boldsymbol{\alpha}$ の i 番目の成分 α^i は $\psi_i(x_i)$ と次の関係がある．

$$\alpha^i = \frac{1}{2}\ln\frac{\psi_i(x_i=+1)}{\psi_i(x_i=-1)}.$$

確率伝搬法の問題は (4) 式の分布から周辺分布を求めることである．

確率変数 x_i の分布はベルヌーイ分布であるから，$q(x_i)$ を表現するにはひとつのパラメータを用いれば十分である．x_i の各成分が独立で一つずつパラメータを持つ確率分布の族 M_0 を考える．

$$\begin{aligned} M_0 &= \big\{p_0(\boldsymbol{x};\boldsymbol{\theta}) \;\big|\; p_0(\boldsymbol{x};\boldsymbol{\theta}) = \exp(c_0(\boldsymbol{x}) + \boldsymbol{\theta}\cdot\boldsymbol{x} - \varphi_0(\boldsymbol{\theta}))\big\} \\ \boldsymbol{\theta} &= (\theta^1,\cdots,\theta^N)^T,\ \boldsymbol{\theta}\in\Re^N,\ \varphi_0(\boldsymbol{\theta}) = \sum_i \ln(e^{\alpha^i+\theta^i} + e^{-(\alpha^i+\theta^i)}). \end{aligned} \tag{5}$$

$\boldsymbol{\theta}$ は自然パラメータと呼ばれ，M_0 のひとつの座標系を与える．M_0 に属する分布は各成分が独立である．逆に x_i が常に -1 または $+1$ をとる特殊な場合を除き，全ての独立な分布は M_0 の分布として表現できる．したがって確率伝搬法の扱う問題は (4) 式で定義された $q(\boldsymbol{x})$ からその周辺分布に対応する $\boldsymbol{\theta}$ 座標を求めることと定式化できる．すなわち $\prod_i q(x_i) = p_0(\boldsymbol{x};\hat{\boldsymbol{\theta}}) \in M_0$ となる $\hat{\boldsymbol{\theta}}$ を求めることが目的となる．ここで次の KL (Kullback-Leibler) 情報量を考える．

$$D(q(\boldsymbol{x}); p_0(\boldsymbol{x};\boldsymbol{\theta})) = \sum_{\boldsymbol{x}} q(\boldsymbol{x})\ln\frac{q(\boldsymbol{x})}{p_0(\boldsymbol{x};\boldsymbol{\theta})}.$$

KL 情報量 $D(q(\boldsymbol{x}); p(\boldsymbol{x}))$ は常に非負で全ての $\boldsymbol{x}$ に対して $q(\boldsymbol{x}) = p(\boldsymbol{x})$ のときのみ 0 となる．すると $\hat{\boldsymbol{\theta}}$ は $D(q(\boldsymbol{x}); p_0(\boldsymbol{x};\boldsymbol{\theta}))$ を最小にすることがわかる．簡単に証明をする．$D(q(\boldsymbol{x}); p_0(\boldsymbol{x};\boldsymbol{\theta}))$ を $\boldsymbol{\theta}$ で微分し 0 とおくと，次式が得られる．

$$\sum_{\boldsymbol{x}} \boldsymbol{x} q(\boldsymbol{x}) = \frac{\partial\varphi_0(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}.$$

簡単な計算から

$$\frac{\partial\varphi_0(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \sum_{\boldsymbol{x}} \boldsymbol{x} p_0(\boldsymbol{x};\boldsymbol{\theta})$$

であるので $D(q(\boldsymbol{x}); p_0(\boldsymbol{x};\boldsymbol{\theta}))$ を最小とする $\hat{\boldsymbol{\theta}}$ は次式を満たす．

$$\sum_{\boldsymbol{x}} \boldsymbol{x} q(\boldsymbol{x}) = \sum_{\boldsymbol{x}} \boldsymbol{x} p_0(\boldsymbol{x};\hat{\boldsymbol{\theta}}).$$

これは各 x_i の期待値が $q(\boldsymbol{x})$ と $p_0(\boldsymbol{x};\hat{\boldsymbol{\theta}})$ とで等しいことを示している．このとき $q(\boldsymbol{x})$ の周辺分布の積 $\prod_i q(x_i)$ と $p_0(\boldsymbol{x};\hat{\boldsymbol{\theta}})$ とは等しい．

以下では他の分野における興味深い問題が (4) 式の $q(\boldsymbol{x})$ の周辺化の問題として定義できることを示す．

2.3 統計物理

まず，統計物理における古典的なスピングラスの問題を考える．隣あったスピンは J_{ij} で示される相互作用を有しているとする．全てのスピンをベクトル $\boldsymbol{x} = (x_1, \cdots, x_N)^T$ で表すと同時分布は

$$q(\boldsymbol{x}) = \frac{1}{Z} \exp\Big(\sum_i h_i x_i + \sum_{i<j} J_{ij} x_i x_j\Big) \tag{6}$$

と定義される．この同時分布から各スピンの周辺分布を求める問題を考える．(6) 式は $c_0(\boldsymbol{x}) = \sum_i h_i x_i$, $c_r(\boldsymbol{x}) = J_{ij} x_i x_j$ とおくと次のように表現できる．

$$q(\boldsymbol{x}) = C \exp\Big(c_0(\boldsymbol{x}) + \sum_r c_r(\boldsymbol{x})\Big).$$

スピングラスの問題を表すグラフにはループが含まれることから正しい結果は期待できないが，確率伝搬法を用いて周辺分布を近似することはできる．そしてその計算手法が統計物理で古くから用いられているベーテ近似や TAP(Thouless-Anderson-Palmer) 方程式と関係のあることが指摘されている [10, 11, 20] ．

近年，計算機の発達に伴い MCMC(Markov chain Monte Carlo) のようなサンプリング手法によりこの問題を解くことも可能となった ([5] などを参照のこと)．しかし確率伝搬法は計算量が少なく，有効な手法であることに変わりない．

2.4 誤り訂正符合

近年，ターボ符号 [3]，LDPC(low-density parity-check: 低密度パリティ検査) 符号 [4] といった高性能な誤り訂正符号が注目されている (詳しくは本号の井坂氏による別稿を参照のこと)．これらの復号アルゴリズムが確率伝搬法と同等の計算手法であることが示されている [14, 15]．ここではこれらの復号の問題も (4) 式の周辺分布を求めることと同値であることを示す．

ターボ復号

情報ブロック $\boldsymbol{x} = (x_1, \cdots, x_N)^T, x_i \in \{-1, +1\}$ を記憶のない BSC(binary symmetric channel: 二元対称通信路) を介して送ることを考える．簡単のため BSC を考えるが，一般の記憶の無い通信路への拡張は簡単である [8]．ターボ符号は畳み込み符号として実装されるが，本稿ではブロック符号として扱う．ターボ符号は一つの符号語に対して 2 つのエンコーダを用いて 2 つのパリティ検査語を作成する．それぞれを $\boldsymbol{y}_1 = (y_{11}, \cdots, y_{1M})^T$, $\boldsymbol{y}_2 = (y_{21}, \cdots, y_{2M})^T$, $y_{1j}, y_{2j} \in \{-1, +1\}$ とする．$\boldsymbol{y}_r$, $r = 1, 2$ は $\boldsymbol{x}$ の関数である．$(\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2)$ を通信路によって送信し，$(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}_1, \tilde{\boldsymbol{y}}_2)$, $\tilde{x}_i, \tilde{y}_{1j}, \tilde{y}_{2j} \in \{-1, +1\}$ と受信されたとする．この受信語に基づき，符号語を推定する．

ターボ復号では，MPM (maximum posterior marginal) 復号の解を求めることを最終目標とする．MPM 復号では $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}_1, \tilde{\boldsymbol{y}}_2)$ の条件付きでの $\boldsymbol{x}$ の分布

$p(\boldsymbol{x}|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}_1,\tilde{\boldsymbol{y}}_2)$ を考え，その分布を $\boldsymbol{x}$ の各成分について周辺化し，周辺化された分布を最大にする符号を復号語とする．まず $p(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}_1,\tilde{\boldsymbol{y}}_2|\boldsymbol{x})$ について考える．通信路が記憶のない BSC であることから

$$p(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}_1,\tilde{\boldsymbol{y}}_2|\boldsymbol{x}) = p(\tilde{\boldsymbol{x}}|\boldsymbol{x})p(\tilde{\boldsymbol{y}}_1|\boldsymbol{x})p(\tilde{\boldsymbol{y}}_2|\boldsymbol{x})$$

$$p(\tilde{\boldsymbol{x}}|\boldsymbol{x}) = \exp\left(\beta\tilde{\boldsymbol{x}}\cdot\boldsymbol{x} - N\psi(\beta)\right),\ \ p(\tilde{\boldsymbol{y}}_r|\boldsymbol{x}) = \exp\left(\beta\tilde{\boldsymbol{y}}\cdot\boldsymbol{y}_r(\boldsymbol{x}) - M\psi(\beta)\right),\ \ r=1,2$$

と書ける．β は正の実数で，通信路のビットの誤り率 σ とは $\sigma = (1-\tanh\beta)/2$ の関係がある．$c_0(\boldsymbol{x}) = \beta\tilde{\boldsymbol{x}}\cdot\boldsymbol{x}$, $c_1(\boldsymbol{x}) = \beta\tilde{\boldsymbol{y}}_1\cdot\boldsymbol{y}_1(\boldsymbol{x})$, $c_2(\boldsymbol{x}) = \beta\tilde{\boldsymbol{y}}_2\cdot\boldsymbol{y}_2(\boldsymbol{x})$ と置くと，$p(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}_1,\tilde{\boldsymbol{y}}_2|\boldsymbol{x})$ は

$$p(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}_1,\tilde{\boldsymbol{y}}_2|\boldsymbol{x}) = \exp\left(c_0(\boldsymbol{x}) + c_1(\boldsymbol{x}) + c_2(\boldsymbol{x}) - (N+2M)\psi(\beta)\right)$$

となる．$\boldsymbol{x}$ の事前分布として一様分布を考えれば事後分布は次のようになる

$$p(\boldsymbol{x}|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}_1,\tilde{\boldsymbol{y}}_2) = \frac{p(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}_1,\tilde{\boldsymbol{y}}_2|\boldsymbol{x})}{\sum_{\boldsymbol{x}} p(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}_1,\tilde{\boldsymbol{y}}_2|\boldsymbol{x})} = C\exp\Big(c_0(\boldsymbol{x}) + \sum_{r=1}^{2} c_r(\boldsymbol{x})\Big). \qquad (7)$$

この分布は (4) 式と同じ形をしている．$\boldsymbol{x}$ の次元は通常数百から数千であるため，$p(\boldsymbol{x}|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}_1,\tilde{\boldsymbol{y}}_2)$ の周辺分布を直接計算することはできない．一方，$c_1(\boldsymbol{x})$, $c_2(\boldsymbol{x})$ のうち片方のみ，すなわち片方のパリティ語のみを用いた復号は BCJR アルゴリズムによって可能となる．そして (7) 式の分布に対し確率伝搬法を用いると，ターボ復号と同じアルゴリズムになることが示されている [15].

LDPC 符号

LDPC 符号の復号も (4) 式によって表される分布の周辺分布を求める問題として定式化できることを示す．図 2 に本稿で扱う LDPC 符号の構造を示す．$\boldsymbol{s} = (s_1,\cdots,s_M)^T$, $s_i \in \{-1,+1\}$, を情報語とする．$C_1 \in \{0,1\}^{K\times M}$, $C_2 \in \{0,1\}^{K\times K}$ は疎行列であり，送信側と受信側で共有されているとする．C_2 は $GF(2)$ 上で逆行列を持つとする．パリティ検査行列は

$$H = (C_1\ C_2), \quad H \in \{0,1\}^{K\times N}, \quad N = M+K$$

で与えられる．符号語は

$$G^T = \begin{pmatrix} E_M \\ C_2^{-1}C_1 \end{pmatrix} \mod 2$$

を用いて $\boldsymbol{u} = G^T\boldsymbol{s} \mod 2$ として作られる．なお E_M は M 次の単位行列である．$\boldsymbol{u}$ の最初の M ビットは $\boldsymbol{s}$ と等しい．通信路は 誤り率 σ の BSC を仮定する．符号語 $\boldsymbol{u}$ は通信路によって $\tilde{\boldsymbol{u}}$ となって観測される．ノイズベクトルを $\boldsymbol{x} = (x_1,\cdots,x_N)^T$, $x_i \in \{0,1\}$ とすればシンドローム $\boldsymbol{y} = (y_1,\cdots,y_K)^T$ と $\boldsymbol{x}$ との間には

$$\boldsymbol{y}(\boldsymbol{x}) = H\tilde{\boldsymbol{u}} = H(\boldsymbol{u}+\boldsymbol{x}) = HG^T\boldsymbol{s} + H\boldsymbol{x} = H\boldsymbol{x} \mod 2$$

という関係が成り立つ．通常は観測された $\tilde{\boldsymbol{y}} = \boldsymbol{y}(\boldsymbol{x})$ を厳密に満たす $\boldsymbol{x}$ を計算するが，ここではさらに別の誤り率 σ' で表される BSC を通した結果が $\tilde{\boldsymbol{y}}$ と

して観測されたという状況を考える．σ' は非常に小さくて構わないので，実際には通常の LDPC 復号と同じ問題だとして良い (確率分布が厳密に正であるという条件を課すためこのような扱いが必要となる)．

以下，各ビットを $\{0,1\}$ の表現から $\{-1,+1\}$ の表現に変える．$0 \to +1$, $1 \to -1$ である．$\boldsymbol{x}$ の条件付での $\tilde{\boldsymbol{y}}$ の分布は以下のように与えられる．

$$p(\tilde{\boldsymbol{y}}|\boldsymbol{x}) = \exp(\rho\tilde{\boldsymbol{y}} \cdot \boldsymbol{y}(\boldsymbol{x}) - K\psi(\rho)) = \exp(c_1(\boldsymbol{x}) + \cdots + c_K(\boldsymbol{x}) - K\psi(\rho)),$$
$$c_r(\boldsymbol{x}) \stackrel{\text{def}}{=} \rho\tilde{y}_r y_r(\boldsymbol{x}), \quad r = 1, \cdots, K.$$

ここで $\sigma' = (1 - \tanh\rho)/2$ である．$\boldsymbol{x}$ の事前分布を $\omega_0(\boldsymbol{x})$ として次のように定義すると，

$$\omega_0(\boldsymbol{x}) = \exp(\beta\mathbf{1}_N \cdot \boldsymbol{x} - N\psi(\beta)) = \exp(c_0(\boldsymbol{x}) - N\psi(\beta))$$
$$c_0(\boldsymbol{x}) \stackrel{\text{def}}{=} \beta\textstyle\sum_{i=1}^N x_i, \quad \mathbf{1}_N = (\underbrace{1, \cdots, 1}_{N})^T$$

$\boldsymbol{x}$ の事後分布は

$$p(\boldsymbol{x}|\tilde{\boldsymbol{y}}) = \frac{p(\tilde{\boldsymbol{y}}|\boldsymbol{x})\omega_0(\boldsymbol{x})}{\sum_{\boldsymbol{x}} p(\tilde{\boldsymbol{y}}|\boldsymbol{x})\omega_0(\boldsymbol{x})} = C\exp\Big(c_0(\boldsymbol{x}) + \sum_{r=1}^K c_r(\boldsymbol{x})\Big) \tag{8}$$

となり，これは (4) 式の $q(\boldsymbol{x})$ と等しい．ここでも $p(\boldsymbol{x}|\tilde{\boldsymbol{y}})$ を直接周辺化するのは難しい．そこでパリティを 1 ビットずつに分けて考え，それぞれの結果をまとめる手法が sum-product アルゴリズムである．このアルゴリズムは (8) 式に示される分布に対する確率伝搬法と等しい．

3　情報幾何に基づく枠組み

ここまで，人工知能，統計物理，誤り訂正符号の興味深い問題が確率伝搬法と関連していることを見た．以下ではこの問題を情報幾何によってどのように定式化し，どのような結果を得たかについて述べる．

3.1　準備

$\boldsymbol{x}$ の確率分布の族 S を考える．これは 2^N 個の要素に対する多項分布全体で構成される多様体である．$(2^N - 1)$ 次元の自由度を持ち，指数型分布族である．

$$S = \Big\{p(\boldsymbol{x})\Big|p(\boldsymbol{x}) > 0, \boldsymbol{x} \in \{-1,+1\}^N, \sum_{\boldsymbol{x}} p(\boldsymbol{x}){=}1\Big\}. \tag{9}$$

次に S に含まれる e–平坦，m–平坦な部分多様体を定義する．

e–平坦: 多様体 $M{\in}S$ は，全ての $q(\boldsymbol{x}), p(\boldsymbol{x}){\in}M$ に対し，次の式で定義される $r(\boldsymbol{x};t)$ が M に含まれるとき，e–平坦である．

$$\ln r(\boldsymbol{x};t) = (1-t)\ln q(\boldsymbol{x}) + t\,\ln p(\boldsymbol{x}) + c(t), \qquad t{\in}\Re.$$

$c(t)$ は $\sum_{\boldsymbol{x}} r(\boldsymbol{x};t) = 1$ と規格化するための関数である．

m–平坦: 多様体 $M{\in}S$ は，全ての $q(\boldsymbol{x}), p(\boldsymbol{x}){\in}M$ に対し，次の式で定義される $r(\boldsymbol{x};t)$ が M に含まれるとき，m–平坦である．

$$r(\boldsymbol{x};t) = (1-t)q(\boldsymbol{x}) + tp(\boldsymbol{x}), \qquad t{\in}[0,1].$$

次に m–射影について定義する．

定義 1. M を S の e–平坦な部分多様体とする．$q(\boldsymbol{x})\in S$ から M への m–射影は， M 上の点で，$q(\boldsymbol{x})$ から M への KL 情報量を最小にする点であり，次のように定義する．

$$\Pi_M\circ q(\boldsymbol{x}) = \underset{p(\boldsymbol{x})\in M}{\operatorname{argmin}} D(q(\boldsymbol{x}); p(\boldsymbol{x})).$$

定理 1. $q(\boldsymbol{x})\in S$ から S の e–平坦な部分多様体 M への m–射影 $\Pi_M\circ q(\boldsymbol{x})$ は *1* 点に定まる *[1].* □

(5) 式に定義した M_0 は定義より指数型分布族であり，指数型分布族は e–平坦であることから e–平坦な部分多様体である [1]．したがって任意の確率分布からの m–射影は一意である．真の分布の周辺化は (4) で定義された確率分布から M_0 への m–射影を求めることと等しい．

3.2 確率伝搬法の情報幾何的表現

本節では確率伝搬法の情報幾何的な表現を与える．天下りとなるが，まず確率伝搬法の考え方を示す．確率伝搬法は $q(\boldsymbol{x})$ を $p_0(\boldsymbol{x};\boldsymbol{\theta})$ によって次のように近似しようというものである．

$$\begin{aligned} q(\boldsymbol{x}) &= \exp(c_0(\boldsymbol{x}) + c_1(\boldsymbol{x}) + \quad \cdots \quad + c_L(\boldsymbol{x}) - \ln C) \\ p_0(\boldsymbol{x};\boldsymbol{\theta}) &= \exp(c_0(\boldsymbol{x}) + \boldsymbol{\xi}_1\cdot\boldsymbol{x} + \quad \cdots \quad + \boldsymbol{\xi}_L\cdot\boldsymbol{x} - \varphi_0(\boldsymbol{\theta})) \end{aligned} \tag{10}$$

ただし，$\boldsymbol{\xi}_r \in \Re^N$, $\boldsymbol{\theta} = \sum_r \boldsymbol{\xi}_r$ である．2 つの節の間の相互作用のみがある場合，i と j を含む r 番目の辺に対する $\boldsymbol{\xi}_r$ と (3) 式に示されるメッセージとの間には次の関係がある．

$$\xi_{r,i} = \frac{1}{2}\ln\frac{m_{ji}(x_i=+1)}{m_{ji}(x_i=-1)}, \quad \xi_{r,j} = \frac{1}{2}\ln\frac{m_{ij}(x_j=+1)}{m_{ij}(x_j=-1)}, \quad \xi_{r,k} = 0 \text{ for } k\neq i,j.$$

(10) 式から，確率伝搬法は $c_r(\boldsymbol{x})$ という (一般に $\boldsymbol{x}$ の非線形な) 関数を $\boldsymbol{\xi}_r\cdot\boldsymbol{x}$ という線形な項で近似していることがわかる．各 $\boldsymbol{\xi}_r$ を求めるために，確率伝搬法では次の分布を用いる．

$$p_r(\boldsymbol{x};\boldsymbol{\zeta}_r) = \exp\left(c_0(\boldsymbol{x}) + c_r(\boldsymbol{x}) + \boldsymbol{\zeta}_r\cdot\boldsymbol{x} - \varphi_r(\boldsymbol{\zeta}_r)\right), \quad r = 1,\cdots,L.$$

この分布は $c_r(\boldsymbol{x})$ を含むが周辺化は可能である．周辺化に必要な計算は (3) 式と同等である．再び天下りだが，以下の分布を考える．

$$\begin{aligned} p_0(\boldsymbol{x};\boldsymbol{\theta}) &= \exp(c_0(\boldsymbol{x}) + \boldsymbol{\xi}_1\cdot\boldsymbol{x} + \quad \boldsymbol{\xi}_2\cdot\boldsymbol{x} + \quad \cdots \quad + \boldsymbol{\xi}_L\cdot\boldsymbol{x} - \varphi_0(\boldsymbol{\theta})) \\ p_1(\boldsymbol{x};\boldsymbol{\zeta}_1) &= \exp(c_0(\boldsymbol{x}) + c_1(\boldsymbol{x}) + \quad \boldsymbol{\xi}_2\cdot\boldsymbol{x} + \quad \cdots \quad + \boldsymbol{\xi}_L\cdot\boldsymbol{x} - \varphi_1(\boldsymbol{\zeta}_1)) \\ p_2(\boldsymbol{x};\boldsymbol{\zeta}_2) &= \exp(c_0(\boldsymbol{x}) + \boldsymbol{\xi}_1\cdot\boldsymbol{x} + \quad c_2(\boldsymbol{x}) + \quad \cdots \quad + \boldsymbol{\xi}_L\cdot\boldsymbol{x} - \varphi_2(\boldsymbol{\zeta}_2)) \\ \vdots & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \vdots \\ p_L(\boldsymbol{x};\boldsymbol{\zeta}_L) &= \exp(c_0(\boldsymbol{x}) + \boldsymbol{\xi}_1\cdot\boldsymbol{x} + \quad \boldsymbol{\xi}_2\cdot\boldsymbol{x} + \quad \cdots \quad + c_L(\boldsymbol{x}) - \varphi_L(\boldsymbol{\zeta}_L)) \end{aligned}$$

ここで $\boldsymbol{\zeta}_r = \boldsymbol{\theta} - \boldsymbol{\xi}_r$ とおいた．確率伝搬法では繰り返しアルゴリズムによって $\boldsymbol{\zeta}_r$ あるいは $\boldsymbol{\xi}_r$ を調整し $\boldsymbol{\theta}$ を求める．収束点を $\boldsymbol{\xi}_r^*$, $\boldsymbol{\zeta}_r^*$, $\boldsymbol{\theta}^*$ とおくと以下の関係

が成り立つ．

$$\sum_{\boldsymbol{x}} \boldsymbol{x} p_0(\boldsymbol{x};\boldsymbol{\theta}^*) = \sum_{\boldsymbol{x}} \boldsymbol{x} p_r(\boldsymbol{x};\boldsymbol{\zeta}_r^*), \quad r = 1, \cdots, L.$$

任意の $r(\boldsymbol{x}) \in S$ から M_0 への m–射影によって求まる座標 $\boldsymbol{\theta}$ を $\pi_{M_0} \circ r(\boldsymbol{x})$ と表すことにする．真の周辺分布の $\boldsymbol{\theta}$ 座標は

$$\pi_{M_0} \circ q(\boldsymbol{x}) = \underset{\boldsymbol{\theta} \in \Re^N}{\operatorname{argmin}} D(q(\boldsymbol{x}); p_0(\boldsymbol{x};\boldsymbol{\theta})),$$

と表される．この π_{M_0} を用いると，確率伝搬法は以下のようにかける．

1. $t = 0$ に対し $\boldsymbol{\zeta}_r^t = \mathbf{o}$ とおき， $t = 1$ とする．
2. $p_r(\boldsymbol{x};\boldsymbol{\theta}_r^t)$ $r = 1, \cdots, L$ から M_0 への射影 $\pi_{M_0} \circ p_r(\boldsymbol{x};\boldsymbol{\theta}_r^t)$ を求め，$\boldsymbol{\xi}_r^{t+1}$ を次のように更新する．

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

3. $\boldsymbol{\theta}^{t+1}$, $\boldsymbol{\zeta}_r^{t+1}$ を以下のように更新する．

$$\boldsymbol{\theta}^{t+1} = \sum_r \boldsymbol{\xi}_r^{t+1}, \quad \boldsymbol{\zeta}_r^{t+1} = \boldsymbol{\theta}^{t+1} - \boldsymbol{\xi}_r^{t+1}.$$

4. $\boldsymbol{\theta}^{t+1}$ が収束しなければ step 2 へ戻る．

上の場合には $\boldsymbol{\zeta}_r$, $\boldsymbol{\theta}$ を一括して更新しているが，逐次的に更新することもできる．更新の順序はアルゴリズムによって異なる．

3.3 この枠組みからわかること

本節では，情報幾何的な枠組みから得られた結果を簡単に述べる．

停留点の性質と近似誤差

次の 2 つの条件が同時に満たされることが，$\boldsymbol{\xi}_r^*$, $\boldsymbol{\zeta}_r^*$, $\boldsymbol{\theta}^*$ が確率伝搬法の停留点となるための必要十分条件である [9].

m–**条件** : $\sum_{\boldsymbol{x}} \boldsymbol{x} p_0(\boldsymbol{x};\boldsymbol{\theta}^*) = \sum_{\boldsymbol{x}} \boldsymbol{x} p_r(\boldsymbol{x};\boldsymbol{\zeta}_r^*), \quad r = 1, \cdots, L.$

e–**条件** : $\boldsymbol{\theta}^* = \sum_r \boldsymbol{\xi}_r^* = \sum_r \boldsymbol{\zeta}_r^* / (L-1)$.

ここで m–平坦な多様体 $M(\boldsymbol{\theta})$, e–平坦な多様体を $E(\boldsymbol{\theta})$ を定義する．

$$M(\boldsymbol{\theta}) = \Big\{ p(\boldsymbol{x}) \;\Big|\; \sum_{\boldsymbol{x}} \boldsymbol{x} p(\boldsymbol{x}) = \sum_{\boldsymbol{x}} \boldsymbol{x} p_0(\boldsymbol{x};\boldsymbol{\theta}) \Big\}$$

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

すると次の定理が得られる [9].

定理 2. 停留点では，$p_0(\boldsymbol{x};\boldsymbol{\theta}^*)$, $p_r(\boldsymbol{x};\boldsymbol{\zeta}_r^*)$, $r = 1, \cdots, L$, が $M(\boldsymbol{\theta}^*)$ に含まれ，$p_0(\boldsymbol{x};\boldsymbol{\theta}^*)$, $p_r(\boldsymbol{x};\boldsymbol{\zeta}_r^*)$, $r = 1, \cdots, L$, $q(\boldsymbol{x})$ が $E(\boldsymbol{\theta}^*)$ に含まれる．

証明. $p_0(\boldsymbol{x};\boldsymbol{\theta}^*)$, $p_r(\boldsymbol{x};\boldsymbol{\zeta}_r^*)$, $r=1,\cdots,L$, が $M(\boldsymbol{\theta}^*)$ と $E(\boldsymbol{\theta}^*)$ に含まれることは定義から明らかである．$q(\boldsymbol{x})$ が $E(\boldsymbol{\theta}^*)$ に含まれることは $t_0=-(L-1)$, $t_1=\cdots=t_L=1$ と置き，$\boldsymbol{\theta}^*=\sum_r \boldsymbol{\zeta}_r^*/(L-1)$ を用いれば確かめられる． □

$q(\boldsymbol{x})$ が $M(\boldsymbol{\theta}^*)$ に含まれれば $p(\boldsymbol{x};\boldsymbol{\theta}^*)$ が真の周辺分布を与える．木のグラフの場合には $q(\boldsymbol{x})$ が $E(\boldsymbol{\theta}^*)$ に含まれれば $M(\boldsymbol{\theta}^*)$ にも含まれることが示せるが [9]，一般には e–平坦性と m–平坦性とは一致しないことから $q(\boldsymbol{x})$ は $M(\boldsymbol{\theta}^*)$ には含まれず，$E(\boldsymbol{\theta}^*)$ のみに含まれる．$E(\boldsymbol{\theta}^*)$ で $M(\boldsymbol{\theta}^*)$ を代用する点が確率伝搬法の近似となる．よく似た構造は他の統計物理の手法にも存在する [12, 17].

また，この 2 つの多様体の差が確率伝搬法で得られた結果と真の周辺分布の差である．我々は摂動法によりこの差を見積もり，[7, 8] に結果を示した．

収束性

確率伝搬法で問題となるのは，主に近似誤差と収束性である．近似誤差の解析は前節に述べた通り摂動法による結果を得た．収束性に関してはこれまで様々な改善アルゴリズムが提案されており，情報幾何の枠組みでこれらのアルゴリズムを統一的に表現すれば比較検討ができ，新たなアルゴリズムに発展する可能性もある．

我々はこれまでターボ符号，LDPC 符号の復号法における局所的な安定性を示し [8]，TRP (tree reparameterization)[19]，CCCP (concave-convex procedure)[21] といったアルゴリズムの解析も行った [9]．アルゴリズムの停留点は m–条件，e–条件を満たすことから，片方の条件を常に満たすようにパラメータを制約した上で他方の条件を同時に満たす点を探すのがわかりやすい．確率伝搬法，TRP といった方法は e–条件を制約とし，m–条件を同時に満たすパラメータを探すアルゴリズムであり，CCCP は 2 重アルゴリズムの内側で m–条件を満たす点を探し，外側で e–条件を同時に満たすようにパラメータを調整していることがわかった [9].

これらのアルゴリズムを理解する上で，関連する損失関数を知ることは重要である．確率伝搬法では，ベーテ自由エネルギーが停留点と関連していることが示されている [10, 11, 20] ．定数の差など多少の違いはあるが，ベーテ自由エネルギーは以下のように定義される．

$$\mathcal{F}(\boldsymbol{\theta},\{\boldsymbol{\zeta}_r\}) = D(p_0(\boldsymbol{x};\boldsymbol{\theta});q(\boldsymbol{x})) - \sum_r D(p_0(\boldsymbol{x};\boldsymbol{\theta});p_r(\boldsymbol{x};\boldsymbol{\zeta}_r)).$$

この関数を $\boldsymbol{\zeta}_r$ で微分すると m–条件が停留条件であり，$\boldsymbol{\theta}$ で微分すると e–条件が停留条件であることがわかる．しかし 2 回微分をとってみるとヘシアン行列は必ずしも正定値ではない．我々は [9] に，ベーテ自由エネルギーと確率伝搬法や関連するアルゴリズムの局所安定性に関する詳しい解析結果を示し，さらに新たな改善アルゴリズムを提案した．

4 まとめ

本稿では，様々な分野で用いられている手法が確率伝搬法と同値であること，さらに情報幾何によって確率伝搬法を解析する枠組みが構築できることを示し

た．ループのあるグラフの周辺分布を求める問題は，計算量の面から厳密解を求めることが難しく，簡便な近似法が実用の面から求められている．確率伝搬法は簡便な手法であるが，本稿で述べたように理論的には未解決な問題も多い．理論的な解明が進み，改善法が提案できれば広い分野での応用が期待できる．

統計物理の分野では，関連する手法として菊池近似，クラスタ変分法といった手法が提案されている．これらの手法は確率伝搬法を一般化したものと理解できる．確率伝搬法に比べ，一般に近似精度は上がるが計算量が増え，収束性の問題は依然として残っている．これらの手法も情報幾何の立場から理解できると考えるが，本稿で示した枠組を拡張する必要がある．今後の課題である．

最後になるが，本稿によって確率伝搬法，また情報幾何に興味を持ってもらえたならば幸いである．

参考文献

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

[2] Bahl, L.R., Cocke, J., Jelinek, F., and Raviv. J., Optimal decoding of linear codes for minimizing symbol error rate. IEEE trans. on Inform. Theory, 20 (1974), 284–287.

[3] Berrou, C., Glavieux, A., and Thitimajshima, P., Near Shannon limit error-correcting coding and decoding: Turbo-codes. In Proc. of IEEE Int. Conf. on Communications, 1993, 1064–1070.

[4] Gallager, R.G., Low density parity check codes. IRE trans. Inform. Theory, IT-8 (1962), 21–28.

[5] 伊庭幸人. ベイズ統計と統計物理岩波書店, 東京, 2003.

[6] 池田思朗, 田中利幸, 甘利俊一. ターボ復号の情報幾何. 電子情報通信学会論文誌, J85-D-II (2002), 758–765.

[7] Ikeda, S., Tanaka, T., and Amari, S., Information geometrical framework for analyzing belief propagation decoder. NIPS 14 (edited by Dietterich, T.G., Becker, S., and Ghahramani, Z.), The MIT Press, 2002, 407–414.

[8] Ikeda, S., Tanaka, T., and Amari, S., Information geometry of turbo codes and low-density parity-check codes. to appear in IEEE trans. on Inform. Theory.

[9] Ikeda, S., Tanaka, T., and Amari, S., Stochastic reasoning, free energy, and information geometry. to appear in Neural Computation.

[10] Kabashima, Y., and Saad, D., Belief propagation vs. TAP for decoding corrupted messages. Europhys. Letters, 44 (1998), 668–674.

[11] Kabashima, Y., and Saad, D., The TAP approach to intensive and extensive connectivity systems. Advanced Mean Field Methods – Theory and Practice(edited by Opper, M., and Saad, D.), The MIT Press, 2001, 65–84.

[12] Kappen, H.J., and Wiegerinck, W.J., Mean field theory for graphical models. Advanced Mean Field Methods – Theory and Practice(edited by Opper, M., and Saad, D.), The MIT Press, 2001, 37–49.

[13] Lauritzen, S.L., and Spiegelhalter, D.J., Local computations with probabilities on graphical structures and their application to expert systems. J. the Royal Stat. Soc. B, 50 (1988), 157–224.

[14] MacKay, D.J.C., Good error-correcting codes based on very sparse matrices. IEEE trans. Inform. Theory, 45 (1999), 399–431.

[15] McEliece, R.J., MacKay, D.J.C., and Cheng, J-F., Turbo decoding as an instance of Pearl's "belief propagation" algorithm. IEEE J. Select. Areas in Commun., 16 (1998), 140–152.

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

[17] Tanaka, T., Information geometry of mean-field approximation. Neural Computation, 12 (2000), 1951–1968.

[18] 田中利幸, 人工知能と確率推論. 数理科学, 489 (2004).

[19] Wainwright, M., Jaakkola, T., and Willsky, A., Tree-based reparameterization for approximate inference on loopy graphs. NIPS 14 (edited by Dietterich, T.G., Becker, S., and Ghahramani, Z.), The MIT Press, 2002, 1001–1008.

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

[21] Yuille, A.L. and Rangarajan, A., The concave-convex procedure. Neural Computation, 15 (2003), 915–936.

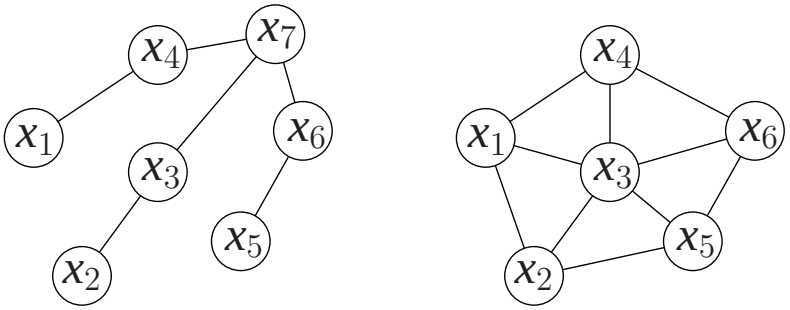

Fig. 1. 木のグラフとループのあるグラフ.

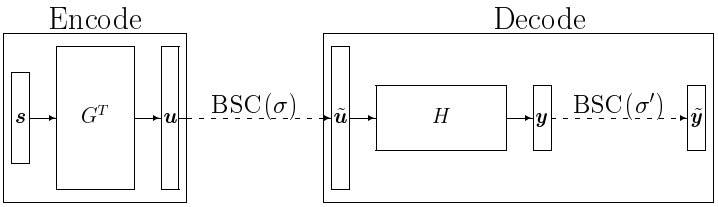

Fig. 2. LDPC 符号.