社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

一般ボルツマンマシンにおける平均場近似自由エネルギーの漸近的挙動

西山　悠†　渡辺　澄夫††

† 東京工業大学総合理工学研究科知能システム科学専攻 〒 226-8503 横浜市緑区長津田 4259
†† 東京工業大学精密工学研究所 〒 226-8503 横浜市緑区長津田 4259
E-mail: †nishiyudesu@cs.pi.titech.ac.jp, ††swatanab@pi.titech.ac.jp

あらまし　人工神経回路網，混合分布，ベイジアンネット等の特異モデルの学習に，ベイズ学習の有効性が示されている．計算困難なベイズ学習に対し，平均場近似を利用したアルゴリズムが用いられ，実問題への有効性が確認されている．近年，平均場近似学習について，汎化誤差，自由エネルギーの理論的な研究がされている．理論的な研究によって，平均場近似のベイズ学習に対する近似精度が明らかとなり，モデル選択への応用の基礎にもつながる．本論文では，一般のボルツマンマシンを考え，特異モデルにおいて一般に縮退するフィッシャー情報行列に対し，零でない固有値の個数を数えることにより，平均場近似学習における自由エネルギーについて，漸近形の上界を理論的に導出する．

キーワード　特異モデル，自由エネルギー，平均場近似，ボルツマンマシン

Asymptotic Behavior of Free Energy of General Boltzmann Machines in Mean Field Approximation

Yu NISHIYAMA† and Sumio WATANABE††

† Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology, 4259 Nagatsuta, Midori-ku, Yokohama, 226-8503 Japan
†† Precision and Intelligence Laboratory, Tokyo Institute of Technology, 4259 Nagatsuta Midori-ku, Yokohama, 226-8503 Japan
E-mail: †nishiyudesu@cs.pi.titech.ac.jp, ††swatanab@pi.titech.ac.jp

Abstract　In the Bayesian learning, which generally requires huge computational costs, the algorithms based on the mean field approximation have shown us the effectiveness in the practical information systems. Recently, the generalization error or free energy in the mean field approximation has been theoretically studied. The theoretical results enable us to know the accuracy of the approximation and contribute to the foundation of a model selection in statistical singular machines. In this paper, we show that the upper bounds of the asymptotic free energies are theoretically obtained by counting the number of non-0 eigenvalues of Fisher information matrices and derive the upper bound in the learning model of general Boltzmann machines.

Key words　Singular Learning Machines, Free Energy, Mean Field Approximation, Boltzmann Machines

1. ま え が き

人工神経回路網，混合分布，ベイジアンネット等の学習モデルは，制御，時系列予測，パターン認識などの情報システムの分野に応用が広がっている．しかしながら，多くの実用的な学習モデルは，数学的な側面から見れば，学習モデルを記述するモデルパラメータと確率分布が 1 対 1 対応とはならない特異モデルであることが知られている．特異モデルでは，真の確率分布を表現する真のパラメータが，1 点ではなく解析的集合の広がりをもつことから，それらの点においてフィッシャー情報行列式が 0 となり，正則なフィッシャー情報行列を持つと仮定した下での統計的漸近論は適用できない．

この問題に対して，代数幾何学的な手法を用いることにより，特異モデルのベイズ汎化誤差の漸近論が理論的に明らかにされ，ベイズ学習の有効性が示されている [1] 〜 [4]．しかしながら，ベイズ学習に現れるベイズ事後分布は，一般に高次元積分を含むことから，ベイズ予測分布などのベイズ事後分布を含む計算は一般に困難である．そこで比較的少ない計算量で実現するた

めの一つの近似方法として平均場近似 (変分ベイズ) が用いられている．学習における平均場近似は，計算が困難なベイズ事後分布を，計算が比較的容易となる，パラメータを互いに独立な分布に近似し，カルバック距離の意味で最もベイズ事後分布に近い分布を選ぶものである．平均場近似を利用したアルゴリズムは，その効率的な計算量から，実問題への有効性が確認されている．

近年，平均場近似を用いた学習において，汎化誤差，自由エネルギー (確率的複雑さ) の理論的な研究がされている．理論的な研究によって，平均場近似アルゴリズムのベイズ学習に対する近似精度を知ることができ，平均場近似で学習した際に，局所解に陥ったかどうかを検証するための基盤ともなる．また，特異モデルにおいて提案されているモデル選択規準 SingIC [5] は，それら理論的に解明された漸近論の情報を利用したものとなっている．現在では，縮小ランク回帰モデル [6]，混合正規分布 [7]，隠れマルコフモデル [8]，確率文脈自由文法 [9]，多層パーセプトロン [10] の学習モデルについて平均場近似自由エネルギーの漸近論が理論的に求められている．

本論文は，特異モデルの 1 つとして，統計物理学と関係の深いボルツマンマシンを対象とする．ボルツマンマシンは，数学的な観点から見れば，情報学で広く用いられているベイジアンネットあるいはグラフィカルモデルの 1 つといえ，統計物理学の磁性体のモデルとも関連する学習モデルである．本論文では，特に，**3.** で与えられる一般的なボルツマンマシンを考え，平均場近似自由エネルギーの漸近形の上界を理論的に導出する．上界を導出する方法としては，ベイズ事後分布を正規分布族に近似した場合 [10] について考えている．

なお，本稿は，3 月 NC 研究会で発表した，完全 2 部グラフ型ボルツマンマシンの場合を拡張し，一般のボルツマンマシンについて，平均場近似自由エネルギーの漸近形の上界を導出したものになっている．

2. 学 習 理 論

2.1 ベイズ学習

ここではベイズ学習の数学的枠組について述べる．一般にデータをベクトル $x \in R^M$ と表し，データが真の確率分布 $q(x)dx$ に従う確率変数 X から独立に得られるものとする．得られた n 個の学習データ $X^n = (X_1, X_2, ..., X_n)$ を使って，真の確率密度関数 $q(x)$ を，条件付き確率密度関数 $p(x|\theta)$ の学習モデルによって推測する．$\theta \in R^d$ は学習モデルの自由度でありモデルパラメータである．学習する前のパラメータの事前分布を $\varphi(\theta)$ とすると，ベイズの定理から，

$$p(\theta|X^n) = \frac{1}{Z(X^n)}\varphi(\theta)\prod_{i=1}^{n} p(X_i|\theta)$$

が成り立つ．ここで $Z(X^n)$ は正規化定数であり，

$$Z(X^n) = \int \varphi(\theta)\prod_{i=1}^{n} p(X_i|\theta)d\theta$$

である．$p(\theta|X^n)$ は，n 個のデータを学習後のパラメータの分布と言え，ベイズ事後分布と呼ばれる．ベイズ事後分布 $p(\theta|X^n)$ を使って，学習後の x の確率分布を

$$p(x|X^n) = \int p(x|\theta)p(\theta|X^n)d\theta$$

と構成する方法をベイズ予測という．$p(x|X^n)$ はベイズ予測分布と呼ばれる．ベイズ予測された確率密度関数 $p(x|X^n)$ と真の確率密度関数 $q(x)$ との関数の近さを表すのに，

$$G(n) = E_{X^n}\{\int q(x)\log\frac{q(x)}{p(x|X^n)}dx\}$$

で定義される量が用いられる．ここで $\{\}$ の中身はカルバック距離であり，学習データ X^n に依存することから，$E_{X^n}\{\}$ によって平均をとっている．$G(n)$ はベイズ学習後の，真の分布との予測誤差を与えていると言え，ベイズ汎化誤差と呼ばれる．

2.2 特異モデルの学習理論

ベイズ事後分布 $p(\theta|X^n)$ は，分母分子を定数 $\prod_{i=1}^{n} q(X_i)$ で等しく割ることで

$$p(\theta|X^n) = \frac{e^{-nH_n(\theta)}\varphi(\theta)}{\bar{Z}(X^n)}$$

と変形できる．ここで $\bar{Z}(X^n)$ は新しい正規化定数であり，$H_n(\theta)$ は

$$H_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\log\frac{q(X_i)}{p(X_i|\theta)}$$

で定義される経験カルバック情報量である．今，

$$\begin{aligned} F(n) &= E_{X^n}\{-\log\bar{Z}(X^n)\} \\ &= E_{X^n}\{-\log\int e^{-nH_n(\theta)}\varphi(\theta)d\theta\} \end{aligned} \quad (1)$$

と定義すると，$F(n)$ はベイズ汎化誤差 $G(n)$ と

$$G(n) = F(n+1) - F(n) \quad (2)$$

の関係をもつことが知られる．このことからベイズ汎化誤差を求めるのに $F(n)$ を計算すればよいことがわかる．式 (1) の $F(n)$ は統計物理学とのアナロジーから自由エネルギー (確率的複雑さ) と呼ばれる．また $\bar{Z}(X^n)$ は分配関数と呼ばれる．

自由エネルギーは，代数幾何学的な方法を用いることで，漸近形 ($n \to \infty$) が求められる．$z \in C$，$H(\theta)$ をカルバック情報量して，関数

$$J(z) = \int H(\theta)^z\varphi(\theta)d\theta$$

を考える．$J(Z)$ はゼータ関数と呼ばれる．この極は負の有理数上にあり，原点に最も近い極を $-\lambda$，その位数を m としたとき，$n \to \infty$ に対して

$$F(n) \sim \lambda\log n - (m-1)\log\log n + O(1) \quad (3)$$

と漸近展開できることが知られる [1]．したがって汎化誤差 $G(n)$ も漸近展開可能であれば，式 (2) の差分の関係式から，汎化誤差 $G(n)$ は，

$$G(n) \sim \frac{\lambda}{n} - \frac{m-1}{\log\log n} \quad (4)$$

と漸近展開できる．

2.3 平均場近似

次に学習における平均場近似について述べる．式 (1) の自由エネルギー $F(n)$ は，Jensen の不等式を用いることで，$f(\theta)$ をパラメータ上の任意試験分布としたもとに，以下の上界を持つ．

$$\begin{aligned} F(n) &= E_{X^n}\{-\log \int f(\theta)\frac{e^{-nH_n(\theta)}\varphi(\theta)}{f(\theta)}d\theta\} \\ &\leqq E_{X^n}\{-\int f(\theta)\log \frac{e^{-nH_n(\theta)}\varphi(\theta)}{f(\theta)}d\theta\} \\ &= E_{X^n}\{\int f(\theta)\log f(\theta)d\theta + n\int f(\theta)\tilde{H}_n(\theta)d\theta\} \end{aligned} \quad (5)$$

ただし $\tilde{H}_n(\theta)$ は，

$$\tilde{H}_n(\theta) = H_n(\theta) - \frac{1}{n}\log\varphi(\theta)$$

である．統計物理学との対応では，式 (5) の第 1 項はエントロピー項であり，第 2 項はエネルギー項である．$f(\theta) = e^{-n\tilde{H}_n(\theta)}/\bar{Z}(X^n)$ のとき等号が成立する．$f(\theta)$ として，特に，すべての変数を互いに独立な

$$\bar{f}(\theta) = \prod_{j=1}^{d} f_j(\theta_j) \quad (6)$$

の関数形に制限し，式 (5) 右辺を最小にする確率分布を用いるとき，近似事後分布 $\bar{f}(\theta)$ を事後分布における平均場近似と呼ぶ．このとき，式 (5) の右辺である自由エネルギーの上界の最小値

$$\bar{F}(n) = E_{X^n}[\min_{\bar{f}(\theta)}\{\int \bar{f}(\theta)\log\bar{f}(\theta)d\theta + n\int \bar{f}(\theta)\tilde{H}_n(\theta)d\theta\}] \quad (7)$$

を平均場近似自由エネルギーと呼ぶ．さらに，式 (7) の $\bar{F}(n)$ は，平均と最小化の順序を考え，Jensen の不等式を用いることで，以下の不等式が成り立つ．

$$\begin{aligned} \bar{F}(n) &\leqq \min_{\tilde{f}(\theta)}\{\int \tilde{f}(\theta)\log\tilde{f}(\theta)d\theta + n\int \tilde{f}(\theta)E_{X^n}[\tilde{H}_n(\theta)]d\theta\} \\ &= \min_{\tilde{f}(\theta)}\{\int \tilde{f}(\theta)\log\tilde{f}(\theta)d\theta + n\int \tilde{f}(\theta)\tilde{H}(\theta)d\theta\} \\ &\equiv \tilde{F}(n) \end{aligned} \quad (8)$$

ここで $\tilde{F}(n)$ の中の $\tilde{H}(\theta)$ は，経験カルバック情報量 $H_n(\theta)$ の平均であるカルバック情報量

$$H(\theta) = \int q(x)\log\frac{q(x)}{p(x|\theta)}dx \quad (9)$$

を用いて

$$\tilde{H}(\theta) = H(\theta) - \frac{1}{n}\log\varphi(\theta) \quad (10)$$

である．カルバック情報量において $H(\theta) = 0$ になることと $q(x) = p(x|\theta)$ になることは必要十分条件である．

3. 学習モデル

3.1 学習モデル

本論文で扱う学習モデルはボルツマンマシンとし，以下で表される一般的なボルツマンマシンを考える．図 1 において，$\{x_j\}_{j=1}^{M}$ を入出力素子，$\{y_i\}_{i=1}^{K}$ を隠れ素子とする．$\{x_j\}_{j=1}^{M}$，$\{y_i\}_{i=1}^{K}$ はそれぞれすべて $\{+1,-1\}$ の 2 値をとるとする．隠れ素子 $\{y_i\}_{i=1}^{K}$ から異なる I 個のノードを選び，入出力素子 $\{x_j\}_{j=1}^{M}$ から J 個の異なるノードを選ぶ．それによってできる結合を $w_{j_1,\cdots,j_J}^{i_1,\cdots,i_I}x_{j_1}\cdots x_{j_J}y_{i_1}\cdots y_{i_I}$ とする (紙面の都合上，$w_{j_1,\cdots,j_J}^{i_1,\cdots,i_I}X_JY_I$ と表記する)．このとき $I+J$ 体相互作用と表す．例えば，図 1 のように，5 体相互作用で，$I=3$,$J=2$ とし，y_2,y_3,y_5,x_2,x_4 を選べば，結合は $w_{24}^{235}x_2x_4y_2y_3y_5$ である (図 1 には，他の結合の例も書いてある)．l_1 体相互作用,$\cdots$,l_L 体相互作用が同時に存在するとき，$(l_1,\cdots,l_L)$ 体相互作用と表す．このとき図 1 において $(l_1,\cdots,l_L)$ 体相互作用を表す確率分布は，**2.** においてモデルパラメータ θ を w と変更したもとで，

$$p(x,y|w) = \frac{1}{Z_A(w)} \exp\{\sum_{k=1}^{L}\sum_{I,J\geqq 0}^{I+J=l_k}\sum_{i_1<\cdots<i_I}\sum_{j_1<\cdots<j_J} w_{j_1,\cdots,j_J}^{i_1,\cdots,i_I}X_JY_I\} \quad (11)$$

と表される．ここで，$\sum_{i_1<\cdots<i_I}$ は隠れ素子 $\{y_i\}_{i=1}^{K}$ から，異なる I 個を選び出すすべての組み合わせによる和である．$\sum_{I,J\geqq 0}^{I+J=l_k}$ は，l_k 体相互作用となる I,J のすべての組み合わせによる和である．$Z_A(w)$ は規格化定数である．学習モデルは，観測できない隠れ素子 $\{y_i\}_{i=1}^{K}$ によって周辺化することで，

$$p(x|w) = \sum_y \frac{1}{Z_A(w)} \exp\{\sum_{k=1}^{L}\sum_{I,J\geqq 0}^{I+J=l_k}\sum_{i_1<\cdots<i_I}\sum_{j_1<\cdots<j_J} w_{j_1,\cdots,j_J}^{i_1,\cdots,i_I}X_JY_I\} \quad (12)$$

と表される．ここで $\sum_y$ は $\sum_{y_1}\cdots\sum_{y_K}$ であり，それぞれの和は $\{1,-1\}$ をとる．

後に使うために，式 (11) を，隠れ素子 y_α に注目して，以下のように変形する．そのために，まず，

$$\sum_{i_1<\cdots<i_I} = \sum_{\alpha} + \sum_{\ni\alpha} \quad (13)$$

と，y_α を中心にして和を分解する．$\sum_\alpha$ は y_α を含む $\binom{K-1}{I-1}$ 通りの選び方による和であり，$\sum_{\ni\alpha}$ は y_α を含まない残りの $\binom{K-1}{I}$ 通りの組み合わせによる和である．このとき式 (11) は，式 (13) を式 (11) に代入することで，以下，

$$\begin{aligned} &p(x,y|w) = \frac{1}{Z_A(w)} \\ &\exp\{\sum_{k=1}^{L}\sum_{I,J\geqq 0}^{I+J=l_k}\sum_{\alpha}\sum_{j_1<\cdots<j_J} w_{j_1,\cdots,j_J}^{i_1,\cdots,i_I}X_JY_I\} \\ &\exp\{\sum_{k=1}^{L}\sum_{I,J\geqq 0}^{I+J=l_k}\sum_{\ni\alpha}\sum_{j_1<\cdots<j_J} w_{j_1,\cdots,j_J}^{i_1,\cdots,i_I}X_JY_I\} \\ &= \frac{Z_{y_\alpha}(w)}{Z_A(w)} \end{aligned}$$

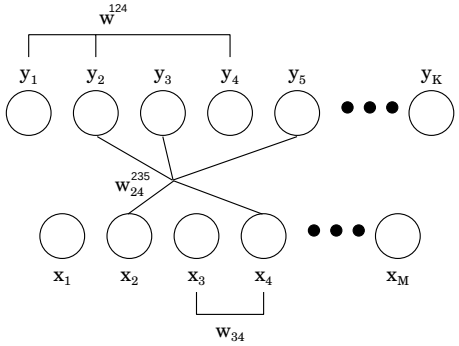

図 1　ボルツマンマシン

$$\exp\{\sum_{k=1}^{L}\sum_{I,J\geqq 0}^{I+J=l_k}\sum_{\alpha}\sum_{j_1<\cdots<j_J} w^{i_1,\cdots,i_I}_{j_1,\cdots,j_J}X_JY_I\}$$
$$\Big/\sum_{y_\alpha}\exp\{\sum_{k=1}^{L}\sum_{I,J\geqq 0}^{I+J=l_k}\sum_{\alpha}\sum_{j_1<\cdots<j_J} w^{i_1,\cdots,i_I}_{j_1,\cdots,j_J}X_JY_I\}$$
$$\equiv \frac{Z_{y_\alpha}(w)}{Z_A(w)}p_\alpha(y_\alpha|y_{\neq\alpha},x,w) \qquad (14)$$

と変形できる．ここで $p_\alpha(y_\alpha|y_{\neq\alpha},x,w)$ は y_α 以外の y,x,w が与えられた下での y_α の条件付き確率を表す．$Z_{y_\alpha}(w)$ は，

$$\sum_{y_\alpha}\exp\{\sum_{k=1}^{L}\sum_{I,J\geqq 0}^{I+J=l_k}\sum_{i_1<\cdots<i_I}\sum_{j_1<\cdots<j_J} w^{i_1,\cdots,i_I}_{j_1,\cdots,j_J}X_JY_I\} \quad (15)$$

と式 (11) の分子の y_α だけによる和である．式 (14) は，確率分布 $p(x,y|w)$ を，y_α を含む項と含まない項に分離したことと言える．

3.2 真の分布

データを発生している真の分布は，式 (12) の形で表現される確率分布とする．真の隠れ変数の個数を K^* （$K^* \leqq K$）とする．このとき，真の確率分布を表現する真の結合パラメータ w^* として，

$$^*w^{i_1,\cdots,i_I}_{j_1,\cdots,j_J} = 0 \quad \text{for} \quad i_I > K^* \qquad (16)$$

を満たすものが存在する．真のパラメータ w^* を用いて，真の分布は $q(x)=p(x|w^*)$ である．

式 (16) を満たす結合パラメータ w^* のとき，式 (14) の p_α は，α が $\alpha\in\{K^*+1,\cdots,K\}$ であれば，式 (16) を p_α に代入して，

$$p_\alpha(y_\alpha|y_{\neq\alpha},x,w^*)=\frac{1}{2} \quad \text{for } \alpha\in\{K^*+1,\cdots,K\} \quad (17)$$

が成り立つ．

4. 主　定　理

ここでは，**3.** で述べたボルツマンマシンについて，式 (8) における平均場近似自由エネルギー $\bar{F}(n)$ の漸近形 $(n\to\infty)$ の上界を計算し，**4.2** の主定理を与える．

4.1 問題設定

主定理の前に，主定理を導くのに用いる，ボルツマンマシンに対する問題設定を述べる．

(i) 式 (8),(10) において，パラメータ w の事前分布 $\varphi(w)$ は，

$$\varphi(w) = (\frac{1}{\sqrt{2\pi}})^d\frac{1}{|\Sigma|^{\frac{1}{2}}}\exp\{-\frac{1}{2}(w-\hat{w})^T\Sigma^{-1}(w-\hat{w})\},$$
$$\Sigma^{\frac{1}{2}} = \mathrm{diag}(\cdots,\sigma^{i_1,\cdots,i_I}_{j_1,\cdots,j_J},\cdots) \qquad (18)$$

の正規分布に従うとする．d はパラメータ次元である．

(ii) 式 (8) において近似事後分布 $\tilde{f}(w)$ は，

$$\tilde{f}(w) = (\frac{1}{\sqrt{2\pi}})^d\frac{1}{|\tilde{\Sigma}|^{\frac{1}{2}}}\exp\{-\frac{1}{2}(w-\hat{w})^T\tilde{\Sigma}^{-1}(w-\hat{w})\},$$
$$\tilde{\Sigma}^{\frac{1}{2}} = \mathrm{diag}(\cdots,\tilde{\sigma}^{i_1,\cdots,i_I}_{j_1,\cdots,j_J},\cdots) \qquad (19)$$

の正規分布の範囲を動くとする．

式 (18)(19) の下で，式 (8) の右辺が，漸近形 $(n\to\infty)$ の主要項の意味で最小になるように，$\hat{w},\tilde{\Sigma}$ を最適化する．

(iii) 式 (9) における真の分布 $q(x)$ は，**3.2** にあるように，学習モデル $p(x|w)$ に含まれる場合を考える．したがって，学習モデルとして $(l_1,\cdots,l_k)$ 体相互作用をもつボルツマンマシンを考えるとき，真の分布も $(l_1,\cdots,l_k)$ 体相互作用をもつボルツマンマシンで表現されるとする．

4.2 主　定　理

[定理 1]　M を入出力素子の個数，K を学習モデルの隠れ素子の個数，$K^*(\leqq K)$ を真の隠れ素子の個数とする．一般に，$(l_1,\cdots,l_L)$ 体相互作用をするボルツマンマシンにおいて，平均場近似自由エネルギーは，以下の上界をもつ．

$$\bar{F}(n)\leqq\sum_{k=1}^{L}\frac{1}{4}\left\{\binom{M+K}{l_k}+\binom{M+K^*}{l_k}\right\}\log n+C. \qquad (20)$$

C は n に依存しない定数である．

[定理 1] とほぼ同じ内容として次が成り立つ．

[定理 2]　M を入出力素子の個数，K を学習モデルの隠れ素子の個数，$K^*(\leqq K)$ を真の隠れ素子の個数とする．一般に，I_k+J_k 体相互作用（$k=1,\cdots,\tilde{L}$）を同時にもつボルツマンマシンにおいて，平均場近似自由エネルギーは，以下の上界をもつ．

$$\bar{F}(n)\leqq\sum_{k=1}^{\tilde{L}}\frac{1}{4}\left\{\binom{M}{I_k}\binom{K}{J_k}+\binom{M}{I_k}\binom{K^*}{J_k}\right\}\log n+C. \qquad (21)$$

C は n に依存しない定数である．

例.1 $<(2,3)$ 体相互作用ボルツマンマシン$>$

2 体相互作用と 3 体相互作用を同時にもつボルツマンマシンは，式 (20) において $L=2$,$l_1=2$,$l_2=3$ から，

$$\bar{F}(n)\leqq\frac{1}{4}\left\{\binom{M+K+1}{3}+\binom{M+K^*+1}{3}\right\}\log n+C$$

である．

例.2 ＜(2) 体相互作用ボルツマンマシン＞

2 体相互作用をもつボルツマンマシンは，式 (20) において $L=1, l_1=2$ として，

$$\bar{F}(n) \leqq \frac{1}{4}\left\{ \binom{M+K}{2} + \binom{M+K^*}{2} \right\} \log n + C$$

である．

例.3 ＜完全 2 部グラフ型ボルツマンマシン＞

完全 2 部グラフ型ボルツマンマシンは，2 体相互作用ボルツマンマシンのなかでも，特に，入出力素子 $\{x_j\}_{j=1}^{M}$，隠れ素子 $\{y_i\}_{i=1}^{K}$ の間にだけ結合パラメータが存在し，入出力素子間同士，隠れ素子間同士には結合が存在しない場合である．したがって，$(I,J)=(1,1)$ の相互作用のみが存在する場合である．このとき，式 (21) において，$\tilde{L}=1, (I,J)=(1,1)$ より，平均場近似自由エネルギー $\bar{F}(n)$ は，

$$\bar{F}(n) \leqq \frac{MK+MK^*}{4} \log n + C \tag{22}$$

の上界をもつ．

この完全 2 部グラフ型ボルツマンマシンにおける平均場近似自由エネルギーの上界は，3 月 NC 研究会で発表したものになっており，これから **4.2** 主定理はそれを拡張したものになっている [11]．

4.3 定理の証明

定理の証明は，以下の補題を利用する．**4.1** の問題設定にあるように，式 (8) の右辺を，近似事後分布 $\tilde{f}(\theta)$ が正規分布の範囲で，最小化するという方法自体は，ボルツマンマシンに限らずとも一般の学習モデルについて，平均場近似自由エネルギー $\bar{F}(n)$ を評価することができる．以下の補題はこれについて述べたものである．一般の学習モデルで成立することから，パラメータは w を用いず，θ を用いている．

[補題] $H(\theta)$ をカルバック情報量とする．$\theta \in R^d$ とする．$H(\hat{\theta})=0$ となる $\hat{\theta}$ で，フィッシャー情報行列 $I(\hat{\theta})$ の対角成分の非零の個数が r 以下となるものが存在するとき，平均場近似自由エネルギー $\bar{F}(n)$ について

$$\bar{F}(n) \leqq \frac{d+r}{4} \log n + C_0$$

が成り立つ．ここで C_0 は n に依存しない定数である．

[補題] の証明

6. 付録に記載.

[定理 1] の証明

「補題」を，**3.** のボルツマンマシンの場合に適用し，定理の証明を得る．補題に従って，式 (12) で表される学習モデルにおいて，真の結合パラメータ w^* におけるフィッシャー情報行列 $I(w^*)$ の対角成分を計算し，非零の個数を数える．

そのために，まず，フィッシャー情報行列の対角成分

$$I^{i_1,\cdots,i_I}_{j_1,\cdots,j_J}(w^*) = \sum_x p(x|w^*) \left(\left. \frac{\partial \log p(x|w)}{\partial w^{i_1,\cdots,i_I}_{j_1,\cdots,j_J}} \right|_{w^*} \right)^2 \tag{23}$$

が 0 となることは，

$$\left. \frac{\partial \log p(x|w)}{\partial w^{i_1,\cdots,i_I}_{j_1,\cdots,j_J}} \right|_{w^*} = 0 \quad \text{for } {}^{\forall}x \tag{24}$$

となることと必要であり十分である．したがって，フィッシャー情報行列のかわりに式 (24) を満たす結合パラメータの個数を調べればよい．また以下では，一般に，

$$\begin{aligned} \left\langle\, f(x,y) \,\right\rangle_{x;p(x,y|w)} &\equiv \sum_x f(x,y) p(x,y|w) \\ \left\langle\, f(x,y) \,\right\rangle_{x,y;p(x,y|w)} &\equiv \sum_x \sum_y f(x,y) p(x,y|w) \end{aligned} \tag{25}$$

と表すことにする．y_α を含む結合パラメータによる式 (24) の左辺は，

$$\begin{aligned} &\left. \frac{\partial \log p(x|w)}{\partial w^{i_1,\cdots,\alpha,\cdots,i_I}_{j_1,\cdots,j_J}} \right|_{w^*} = \frac{1}{p(x|w)} \left. \frac{\partial p(x|w)}{\partial w^{i_1,\cdots,\alpha,\cdots,i_I}_{j_1,\cdots,j_J}} \right|_{w^*} \\ &= -\left\langle\, y_{i_1} \cdots y_\alpha \cdots y_{i_I} x_{j_1} \cdots x_{j_J} \,\right\rangle_{x,y;p(x,y|w^*)} \\ &\quad + \frac{x_{j_1} \cdots x_{j_J}}{p(x|w^*)} \left\langle\, y_{i_1} \cdots y_\alpha \cdots y_{i_I} \,\right\rangle_{y;p(x,y|w^*)} \\ &= -\left\langle\, y_{i_1} \cdots \left\langle\, y_\alpha \,\right\rangle_{y_\alpha;p_\alpha} \cdots y_{i_I} x_{j_1} \cdots x_{j_J} \,\right\rangle_{x,y_{\neq\alpha};Z_{y_\alpha}/Z_A} \\ &\quad + \frac{x_{j_1} \cdots x_{j_J}}{p(x|w^*)} \left\langle\, y_{i_1} \cdots \left\langle\, y_\alpha \,\right\rangle_{y_\alpha;p_\alpha} \cdots y_{i_I} \,\right\rangle_{y_{\neq\alpha};Z_{y_\alpha}/Z_A} \end{aligned} \tag{26}$$

と計算できる．ここで，3 番目の等式は確率分布 $p(x,y|w)$ が式 (14) で表されることを使った．式 (26) において，α として $\alpha \in \{K^*+1, \cdots, K\}$ のとき，式 (17) から，

$$\left\langle\, y_\alpha \,\right\rangle_{y_\alpha;p_\alpha} = \sum_{y_\alpha} y_\alpha \frac{1}{2} = 0 \tag{27}$$

より，式 (26) は 0 となる．このことから，

$$\left. \frac{\partial \log p(x|w)}{\partial w^{i_1,\cdots,\alpha,\cdots,i_I}_{j_1,\cdots,j_J}} \right|_{w^*} = 0 \quad \text{for } \alpha \in \{K^*+1, \cdots, K\} \tag{28}$$

が成り立ち，式 (23) のフィッシャー情報行列の対角成分に対して

$$I^{i_1,\cdots,\alpha,\cdots,i_I}_{j_1,\cdots,j_J}(w^*) = 0 \quad \text{for } \alpha \in \{K^*+1, \cdots, K\} \tag{29}$$

が従う．

l_k 体相互作用のとき，全結合パラメータ数は，$M+K$ 個から l_k 個を選び出す $\binom{M+K}{l_k}$ 個である．フィッシャー情報行列の対角成分が 0 とはならない個数は，式 (29) より $\binom{M+K^*}{l_k}$ 個以下である．$(l_1, \cdots, l_L)$ 体相互作用すべての結合パラメータを足しあげれば，[補題] における d と r の値は，

$$d = \sum_{l_k=1}^{L} \binom{M+K}{l_k}, \qquad r = \sum_{l_k=1}^{L} \binom{M+K^*}{l_k} \tag{30}$$

となる．この d と r を [補題] の式に代入して定理の式が導出される (証明終了).

式 (28) からフィッシャー情報行列はランク落ちし，固有値の非零の個数も r 以下となる．

「定理 2」の証明

「定理 1」の証明とほぼ同様であり略.

5. ま と め

本稿は，一般のボルツマンマシンに対して，ベイズ事後分布を正規分布と近似することで，平均場近似自由エネルギーの漸近形の上界を与えた．**4.** の主定理の結果から直観的にいえることは，一般的なボルツマンマシンは，いずれであっても，真の分布に対して学習モデルのパラメータが冗長であるものに対し，フィッシャー情報行列の成分が 0 となり，0 となる分だけ自由エネルギーの漸近形の主要項が正則モデルの漸近論より小さくなるということである．今後の課題としては，平均場近似自由エネルギーの下界の導出，さらに，ベイズ事後分布を平易な正規分布に近似した下でアルゴリズムを作り，実験と理論の比較がある．

6. 付　　録

（[補題] の証明）

式 (8) の右辺において，近似事後分布 $\tilde{f}(\theta)$ を

$$\tilde{f}(\theta) = (\frac{1}{\sqrt{2\pi}})^d \frac{1}{|\tilde{\Sigma}|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\theta-\hat{\theta})^T\tilde{\Sigma}^{-1}(\theta-\hat{\theta})\}$$
$$\tilde{\Sigma} = \mathrm{diag}(\tilde{\sigma}_1^2,\cdots,\tilde{\sigma}_d^2) \quad (31)$$

の正規分布族に限定する．このとき式 (8) 右辺の第 1 項目は，ガウス分布のエントロピーの計算から，

$$\int \tilde{f}(\theta)\log\tilde{f}(\theta)d\theta = -\frac{d}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{d}\log\tilde{\sigma}_i^2 - \frac{d}{2} \quad (32)$$

と計算できる．式 (32) を式 (8) に代入して，

$$\bar{F}(n) \leqq -\frac{1}{2}\sum_{i=1}^{d}\log\tilde{\sigma}_i^2 + n\int\tilde{f}(\theta)\tilde{H}(\theta)d\theta + C_1 \quad (33)$$

である．ここで事前分布 $\varphi(\theta)$ を正規分布

$$\varphi(\theta) = (\frac{1}{\sqrt{2\pi}})^d \frac{1}{|\Sigma|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\theta-\hat{\theta})^T\Sigma^{-1}(\theta-\hat{\theta})\}$$
$$\Sigma = \mathrm{diag}(\sigma_1^2,\cdots,\sigma_d^2) \quad (34)$$

とすれば，式 (10) から，式 (33) は，

$$\bar{F}(n) \leqq -\frac{1}{2}\sum_{i=1}^{d}\log\tilde{\sigma}_i^2 + \frac{1}{2}\sum_{i=1}^{d}\frac{\tilde{\sigma_i}^2}{\sigma_i^2} + n\int\tilde{f}(\theta)H(\theta)d\theta + C_2 \quad (35)$$

となる．式 (35) 右辺の第 3 項について，$H(\theta)$ を $\hat{\theta}$ のまわりでテーラー展開すれば，

$$n\int\tilde{f}(\theta)\{\sum_{k=0}^{\infty}\frac{1}{k!}\frac{\partial^k H(\theta)}{\partial\theta^k}\Big|_{\hat{\theta}}(\theta-\hat{\theta})^k\}d\theta$$
$$= n\sum_{k=0}^{\infty}\frac{1}{k!}\frac{\partial^k H(\theta)}{\partial\theta^k}\Big|_{\hat{\theta}}\int\tilde{f}(\theta)(\theta-\hat{\theta})^k d\theta$$
$$= nH(\hat{\theta}) + \frac{n}{2!}\sum_{i=1}^{d}I_{ii}(\hat{\theta})\tilde{\sigma}_i^2 + \frac{n}{4!}\sum_{i=1}^{d}\frac{\partial^4 H(\theta)}{\partial\theta_i^4}\Big|_{\hat{\theta}}\tilde{\sigma}_i^2\cdot 3\tilde{\sigma}_i^2$$
$$+\frac{n}{4!}\underbrace{\sum_{j=1}^{d}\sum_{i=1}^{d}}_{j\neq i}\frac{\partial^4 H(\theta)}{\partial\theta_i^2\partial\theta_j^2}\Big|_{\hat{\theta}}\tilde{\sigma}_i^2\tilde{\sigma}_j^2 + \text{高次項} \quad (36)$$

と漸近展開できる．ここで，$I_{ii}(\hat{\theta})$ は，フィッシャー情報行列 $I(\hat{\theta})$ の (i,i) 成分であり，カルバック情報量の二階微分と，フィッシャー情報行列が等しくなる関係，

$$\frac{\partial^2 H(\theta)}{\partial\theta_i\partial\theta_j}\Big|_{\hat{\theta}} = -\frac{\partial^2}{\partial\theta_i\partial\theta_j}\int p(x|\theta^*)\log p(x|\theta)dx\Big|_{\hat{\theta}}$$
$$= \int\frac{1}{p(x|\theta)}\frac{\partial p(x|\theta)}{\partial\theta_i}\frac{\partial p(x|\theta)}{\partial\theta_j}dx\Big|_{\hat{\theta}} = I_{ij}(\hat{\theta}) \quad (37)$$

を用いた．式 (36) において，フィッシャー情報行列に対する仮定から，

$$\tilde{\sigma}_i^2 = \frac{1}{n} \quad \text{for } \{\, i\,;\, I_{ii}(\hat{\theta}) \neq 0\}$$
$$\tilde{\sigma}_i^2 = \frac{1}{n^{\frac{1}{2}}} \quad \text{for } \{\, i\,;\, I_{ii}(\hat{\theta}) = 0\} \quad (38)$$

と設定すれば，$nH(\hat{\theta}) = 0$ が成り立つことも考慮して，式 (36) は，定数オーダとなる．このとき式 (35) は，

$$\bar{F}(n) \leqq \frac{1}{2}\sum_{j=1}^{r}\log n + \frac{1}{4}\sum_{j=r+1}^{d}\log n + C_0$$
$$= \frac{d+r}{4}\log n + C_0$$

となり，補題の式が導出される ([補題] の証明終了).

文　　献

[1] S. Watanabe, “Algebraic analysis for nonidentifiable learning machines,” Neural Computation, vol.13, no.4, pp.899-933, 2001.

[2] M. Aoyagi and S. Watanabe, “Stochastic complexities of Reduced Rank Regression in Bayesian Estimation,” Int. J. Neural Netw., vol.18, no.7, pp.924-933, 2005.

[3] K. Yamazaki, and S. Watanabe, “Singularities in mixture models and upper bounds of stochastic complexitiy,” Int. J. of Neural Netw., vol.16, no.7, pp.1029-1038, 2003.

[4] K. Yamazaki, and S. Watanabe, “Stochastic complexity of hidden Markov models,” Proc. of NNSP, pp.179-188, 2003.

[5] 山崎啓介, 永田賢二, 渡辺澄夫, “特異モデルにおけるモデル選択法の提案,” 信学技報, NC2005-31, July, 2005.

[6] S. Nakajima, and S. Watanabe, “Generalization Error of Variational Bayes Approach in Reduced Rank Regression,” 信学技報, NC2003-213, March, 2005.

[7] K. Watanabe, and S. Watanabe, “Lower bounds of stochastic complexities in variational Bayes learning of gaussian mixture models,” Proc. IEEE conference on Cybernetics and Intelligent Systems, pp.99-104, 2004.

[8] 星野力, 渡辺一帆, 渡辺澄夫, “隠れマルコフモデルの変分ベイズ推定における確率的複雑さについて” 信学技報, NC2004-225, March, 2005.

[9] 星野力, 渡辺一帆, 渡辺澄夫, “確率文脈自由文法の変分ベイズ推定における確率的複雑さについて,” 信学技報, NC2005-49, October, 2005.

[10] 中野修弘, 渡辺澄夫, “ベイズ事後分布実現における平均場近似の精度評価,” 信学技報, NC2004-212, March, 2005.

[11] 西山悠，渡辺澄夫，” 完全 2 部グラフ型ボルツマンマシンにおける平均場近似自由エネルギーの漸近的挙動,” 信学技報, NC2005-172, March, 2006.