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.

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

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　特異モデルの学習においてベイズ学習の有効性が示されている．しかしながらベイズ事後分布の実現は難しい．ベイズ事後分布を少ない計算量で実現する近似方法として，統計物理学で知られる平均場近似が用いられる．平均場近似を利用したアルゴリズムは実問題への有効性が確認されている．近年，平均場近似の近似精度について理論的な研究がされている．理論的な研究によって，正則モデルとの比較を可能にし，モデル選択への応用の基礎にもつながる．本論文では完全 2 部グラフ型ボルツマンマシンにおいて，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．

キーワード　特異モデル，確率的複雑さ，平均場近似，ボルツマンマシン

Stochastic Complexity of Complete Bipartite Graph-type Boltzmann Machines in Mean Field Approximation

Yu NISHIYAMA† and Sumio WATANABE††

† Department of Computational Intelligence and System 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 learning of singular learning machines, the superiority of Bayesian learning is shown. However, it requires huge computational costs to realize the Bayesian a posteriori distribution. To overcome this problem, the mean field approximation, which is originally known in statistical physics, is used in the practical information systems. Recently, the theoretical properties such as generalization error or free energy in the mean field approximation has been studied. The theoretical results give us the comparison with the regular statistical model and the foundation of a model selection. In this paper, we treat the complete bipartite Boltzmann machines and derive the upper bound of asymptotic free energy of the mean field approximation.

Key words　Singular Learning Machines, Stochastic Complexity, Mean Field Approximation, Boltzmann Machines

1. ま え が き

神経回路網，混合分布，ベイジアンネットといった確率を利用した学習モデルは制御，時系列予測，パターン認識などの現実的な情報システムの分野で応用が広まっている．しかしながら，多くの実用的な学習モデルは，数学的見地から見ると，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 対応では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，従来知られた正則なフィッシャー情報行列を仮定したもとでの統計的漸近論が適用できない．

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って，いくつかの特異モデルにおいてベイズ学習の際の汎化誤差の漸近論が理論的に明らかにされ，分布推定であるベイズ学習の有効性が示されている [1] [2] [3] [4]．しかしながら，ベイズ学習の際に現れるベイズ事後分布は，高次元積分を含み，その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) が使われている．平均場近似は，統計物理学で知られている近似手法で，ボルツマン分布のハミルトニアンにおいて，あるパラメータを平均量におきかえることで相互作用がない系に近似する方法である．その結果，個々のパラメータが互いに独立とでき，計算が容易となるものである．学習における平均場近似では，計算が困難なベイズ事後分布を，計算が容易となるパラメータが互いに独立な試験分布に限定し，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に近似するものである．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，その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) について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学習精度を知ることができ，実際に平均場近似で学習した際に，局所解に陥ったかなどの問題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では，それら理論的に解明された数理的な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 [5]，混合正規分布 [6]，隠れマルコフモデル [7]，確率文脈自由文法 [8]，多層パーセプトロン [9] 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形が理論的に求められている．

本論文では統計物理学と関係の深いボルツマンマシンを対象とし，特に数学的な平易さから，完全二部グラフ型のボルツマンマシンについて考え，平均場近似自由エネルギーの学習サンプル数を増やしたときの漸近論について，そ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，ベイズ事後分布を近似する試験分布として正規分布族に限定した場合 [9] について考えている．その結果，平易な正規分布族でさえ，得られた平均場自由エネルギーの漸近形の上界が，従来知られた統計的正則モデルの場合よりも良い振舞いを持つことが示される．

ボルツマンマシンは，古くから歴史のある学習モデルであり，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またはグラフィカルモデルと考えられる．また，その数学的構造は統計物理学の磁性体のモデルとも関係している．

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 個の学習データを学習後のパラメータ分布と言え，ベイズ事後分布と呼ばれる．このベイズ事後分布 $p(\theta|X^n)$ 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，

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

で定義される量を用いる．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，

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

と定義すると，$F(n)$ はベイズ汎化誤差 $G(n)$ と

$$G(n) = F(n+1) - F(n) \tag{2}$$

の関係をもつことが知られる [10] [11]．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．(1) 式の $F(n)$ 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．

ベイズ汎化誤差は，代数幾何学的な方法をつかうことで漸近形が求められる．$z \in C$ とするゼータ関数

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

を考える．ここで $H(\theta)$ はカルバック情報量である．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

$$F(n) \sim \lambda\log n - (m-1)\log\log n + O(1) \tag{3}$$

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，(2) 式の関係から，

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

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．(1) 式，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．

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

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，(5) 式第 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．$f(\theta) = e^{-n\tilde{H}_n(\theta)}$ 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

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

に制限し，自由エネルギーの上界を最小にする試験分布を用いるとき，$\bar{f}(\theta)$ を平均場近似と呼ぶ．このときの (5) 式，自由エネルギーの上界の最小値

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

を平均場近似自由エネルギーと呼ぶ．さらに，平均と最小化の順序を考え，Jensen の不等式を用いると $\bar{F}(n)$ は以下の上界を持つ．

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

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

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

を用いて

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

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図 1 で表される，グラフ理論の言葉で完全 2 部グラフ型のボルツマンマシンである．ここで $\{x_i\}_{i=1}^M$ は入出力素子であり，$\{y_i\}_{i=1}^K$ は隠れ素子である．$\{x_i\}_{i=1}^M$，$\{y_i\}_{i=1}^K$ はそれぞれすべて $\{+1,-1\}$ の 2 値をとり．x_j，y_i 間で w_{ij} の重みで双線形結合しているとする．このとき完全 2 部グラフ型ボルツマンマシンを表す学習モデルは，モデルパラメータ θ を $w = \{w_{ij}\}$ としたもとで，観測できない隠れ素子 y について周辺化し

$$p(x|w) = \frac{\sum_y \exp(\sum_{i=1}^K\sum_{j=1}^M w_{ij}x_jy_i)}{\sum_x\sum_y \exp(\sum_{i=1}^K\sum_{j=1}^M w_{ij}x_jy_i)}$$

である．ここで $\sum_y$ は，$\sum_{y_1}\sum_{y_2}\sum_{y_3}\cdots\sum_{y_K}$ を意味し，それぞれの和は $\{+1,-1\}$ をとるとする．$\sum_x$ についても同様である．$p(x|w)$ を変形すると

$$\begin{aligned} p(x|w) &= \frac{\prod_{i=1}^K\sum_{y_i}\exp(\sum_{j=1}^M w_{ij}x_jy_i)}{Z(w)} \\ &= \frac{\prod_{i=1}^K\cosh(\sum_{j=1}^M w_{ij}x_j)}{Z(w)} \end{aligned} \tag{10}$$

となる．ここで $Z(w)$ は正規化定数である．

$p(x|w)$ は，入出力素子 $\{x_i\}_{i=1}^M$ が $\{+1,-1\}$ の 2 値をとることから離散分布であり，入出力全通り数は 2^M 通りである．よって，$p(x|w)$ を記述するパラメータは 2^M-1 個で十分であり，隠れ素子数 K は，$KM \leqq 2^M - 1$ を満たすものとする．

また $M = 1$ のとき (10) 式は，$\cosh x$ が偶関数であることを考慮すれば

$$p(x|w) = \frac{\prod_{i=1}^K\cosh(w_{i1})}{Z(w)} = \frac{1}{2^M}$$

となり，$p(x|w)$ はモデルパラメータ $\{w\}$ に依存しない．したがって $M \geqq 2$ とする．以下ではこれら 2 つの条件が成立する範囲で考察を行なう．

3.2 真の分布

サンプルを発生している真の分布は K^* （$K^* \leqq K$）個の隠れ素子を持つとする．すなわち，真の結合パラメータ $\{w^*\}$ は

$$\begin{aligned} w_{ij}^* &\neq 0 \quad for\ i \in \{1,2,\cdots,K^*\} \\ w_{ij}^* &= 0 \quad for\ i \in \{K^*+1,\cdots,K\} \end{aligned}$$

であり，このとき $p(x|w^*)$ は，

$$p(x|w^*) = \frac{\prod_{i=1}^{K^*}\cosh(\sum_{j=1}^M w_{ij}^*x_j)}{Z(w^*)} \tag{11}$$

である．

4. 主 定 理

ここでは，本論文の主定理を述べる．

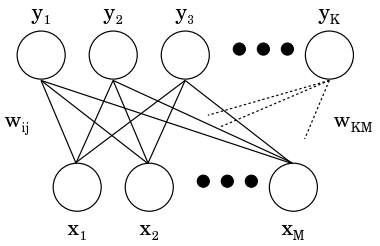

図 1　完全 2 部グラフ型ボルツマンマシン

4.1 問題設定

はじめに「4.2 主定理」を導くのに用いる問題設定を述べる．パラメータの事前分布 $\varphi(w)$ を，

$$\varphi(w) = \frac{1}{Z(\sigma_1)} \exp\{-\frac{\sum_{i=1}^{K}\sum_{j=1}^{M}(w_{ij}-\hat{w}_{ij})^2}{2\sigma_1^2}\} \qquad (12)$$

の正規分布とする．事前分布が正規分布でない場合については，後に「5.1」で述べる．(6) 式に対応する平均場近似する確率分布 $\bar{f}(w)$ を

$$\bar{f}(w) = \frac{1}{Z(L)} \exp\{-\sum_{i=1}^{K}\sum_{j=1}^{M} L_{ij}(w_{ij}-\hat{w}_{ij})^2\} \qquad (13)$$

の正規分布とする．一般に (7),(8) 式における最小化は任意確率分布 $\bar{f}_{ij}(w_{ij})$ の範囲で最小化しなければならないが，確率分布族を正規分布族に制限することで (7),(8) 式の上界を得る．本論文では (8) 式の $\tilde{F}(n)$ の上界を与えることで「4.2 主定理」にある平均場近似の漸近形の結果を得る．$\{L_{ij}\}$ と $\{\hat{w}_{ij}\}$ が，平均場近似自由エネルギーを最小化するのに最適化される変数である．

4.2 主　定　理

[定理]　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^* を真の分布の隠れ素子の個数とする．「3.1 学習モデル」内の 2 つの条件を満たすとする．完全 2 部グラフ型ボルツマンマシンの平均場近似自由エネルギー $\bar{F}(n)$ は次の上界を持つ．

$$\bar{F}(n) \leqq \frac{KM + K^*M}{4}\log n + C.$$

ここで C は n に依存しない定数である．

4.3 定理の証明

はじめに以下の補題を記す．この補題は一般の学習モデルで成立し，パラメータは θ を用いる．

[補題]$\theta \in R^d$ とする．カルバック情報量 $H(\theta)$ において，$H(\hat{\theta}) = 0$ を満たす $\hat{\theta}$ で，集合

$$\{i\,;\ \frac{\partial^2 H(\theta)}{\partial\theta_i^2}\Big|_{\hat{\theta}} \neq 0\}$$

の要素の個数が r 個以下に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

$$\bar{F}(n) \leqq \frac{r+d}{4}\log n + O(1)$$

が成立する．ここで $O(1)$ は n に依存しない定数である．

([補題] の証明)．(4) 式の $\bar{f}(\theta)$ を

$$f_0(\theta) = \frac{\exp\{-\sum_{i=1}^{d} N_i(\theta_i-\hat{\theta}_i)^2\}}{Z(N)} \qquad (14)$$

の正規分布族に限定する．このとき (14) 式を (8) 式に代入して，

$$\bar{F}(n) \leqq \int f_0(\theta)\log f_0(\theta)d\theta + n\int f_0(\theta)\tilde{H}(\theta)d\theta \qquad (15)$$

が成り立つ．(15) 式第一項はガウス分布のエントロピーの計算から，

$$\begin{aligned}\int f_0(\theta)\log f_0(\theta)d\theta &= -\log\prod_{i=1}^{d}(\frac{\pi}{N_i})^{\frac{1}{2}} - \sum_{i=1}^{d} N_i\frac{1}{2N_i} \\ &= \frac{1}{2}\sum_{i=1}^{d}\log N_i - \frac{d}{2}\log\pi - \frac{d}{2} \qquad (16)\end{aligned}$$

である．(16) 式を (15) 式に代入して

$$\bar{F}(n) \leqq \frac{1}{2}\sum_{i=1}^{d}\log N_i + n\int f_0(\theta)\tilde{H}(\theta)d\theta + C_1 \qquad (17)$$

である．ここで事前分布 $\varphi(\theta)$ を正規分布

$$\varphi_0(\theta) = \frac{1}{Z(\sigma)}\exp\{-\frac{\sum_{i=1}^{d}(\theta_i-\hat{\theta}_i)^2}{2\sigma^2}\} \qquad (18)$$

として (17) 式に代入すると，

$$\begin{aligned}\bar{F}(n) \leqq\ & \frac{1}{2}\sum_{i=1}^{d}\log N_i + \frac{1}{2\sigma^2}\sum_{i=1}^{d}\frac{1}{2N_i} \\ & + n\int f_0(\theta)H(\theta)d\theta + C_2 \qquad (19)\end{aligned}$$

となる．以下パラメータ $\{N_i\}$ を主要項の係数が最小になるように最適化する．(19) 式の第三項について $H(\theta)$ を $\hat{\theta}$ の周りでテーラー展開すれば，

$$\begin{aligned}& n\int f_0(\theta)\{\sum_{k=0}^{\infty}\frac{1}{k!}\frac{\partial^k H(\theta)}{\partial\theta^k}\Big|_{\hat{\theta}}(\theta-\hat{\theta})^k\}d\theta \\ =\ & n\sum_{k=0}^{\infty}\frac{1}{k!}\frac{\partial^k H(\theta)}{\partial\theta^k}\Big|_{\hat{\theta}}\int f_0(\theta)(\theta-\hat{\theta})^k d\theta \\ =\ & nH(\hat{\theta}) + \frac{n}{2!}\sum_{i=1}^{d}\frac{\partial^2 H(\theta)}{\partial\theta_i^2}\Big|_{\hat{\theta}}\frac{1}{2N_i} \\ & + \frac{n}{4!}\sum_{i=1}^{d}\frac{\partial^4 H(\theta)}{\partial\theta_i^4}\Big|_{\hat{\theta}}\frac{1}{2N_i}\frac{3}{2N_i} \\ & + \frac{n}{4!}\underbrace{\sum_{j=1}^{d}\sum_{i=1}^{d}}_{j\neq i}\frac{\partial^4 H(\theta)}{\partial\theta_i^2\partial\theta_j^2}\Big|_{\hat{\theta}}(\frac{1}{2N_i})(\frac{1}{2N_j}) \\ & + \text{高次項} \qquad (20)\end{aligned}$$

と漸近展開できる．そこで

$$N_i = n \quad for\ \{\,i\,;\ \frac{\partial^2 H(\theta)}{\partial\theta_i^2}\Big|_{\hat{\theta}} \neq 0\}$$

$$N_i = n^{\frac{1}{2}} \quad for \ \{ i ; \frac{\partial^2 H(\theta)}{\partial \theta_i^2}\Big|_{\hat{\theta}} = 0\} \tag{21}$$

と設定し，$nH(\hat{\theta}) = 0$ であることを考慮すれば，(20) 式は定数オーダーとなる．このとき (19) 式は

$$\begin{aligned}\bar{F}(n) &\leqq \frac{1}{2}\sum_{j=1}^{r}\log n + \frac{1}{4}\sum_{j=r+1}^{d}\log n + O(1)\\ &= \frac{d+r}{4}\log n + O(1)\end{aligned}$$

となり，補題が証明される．([補題] の証明終了)

この [補題] は，一般の学習モデルにおいて成立し，学習モデルの二階微分の計算 (等しくフィッシャー情報量の計算) だけで，補題の平均場近似自由エネルギーの上界が得られることを意味する．この補題を完全 2 部グラフ型ボルツマンマシンの場合に適用して定理を証明する．

([定理] の証明) 完全 2 部グラフ型ボルツマンマシンのカルバック情報量 $H(w)$ は (10),(11) 式を (9) 式に代入することで得られる．$H(w)$ の一階微分，二階微分は，一般に $(\cosh^2 x)^{-1} = 1 - \tanh^2 x$ が成立することを考慮すれば，

$$\begin{aligned}\frac{\partial H(w)}{\partial w_{\alpha\beta}}\Big|_{\hat{w}} &= \left\langle t_{\alpha\beta} \right\rangle_{\hat{w}} - \left\langle t_{\alpha\beta} \right\rangle_{w_*}\\ \frac{\partial^2 H(w)}{\partial w_{\alpha\beta}^2}\Big|_{\hat{w}} &= \left\langle t_{\alpha\beta}^2 \right\rangle_{w^*} - \left\langle t_{\alpha\beta} \right\rangle_{\hat{w}}^2\end{aligned} \tag{22}$$

となる．ここで $t_{\alpha\beta}$ を

$$t_{\alpha\beta} = \tanh(\sum_{j=1}^{M} w_{\alpha j} x_j) x_\beta$$

とおいた．また平均値 $\left\langle f(x, w_1) \right\rangle_{w_2}$ を

$$\left\langle f(x, w_1) \right\rangle_{w_2} = \sum_x f(x, w_1)\, p(x|w_2)$$

とおいた．$\hat{w}$ として $H(\hat{w}) = 0$ を満たす $\hat{w}$ を選ぶとき，$p(x|w^*) = p(x|\hat{w})$ が成立し，$\left\langle \ \right\rangle_{w^*} = \left\langle \ \right\rangle_{\hat{w}}$ が言えることから，

$$\frac{\partial^2 H(w)}{\partial w_{\alpha\beta}^2}\Big|_{\hat{w}} = \left\langle (t_{\alpha\beta} - \left\langle t_{\alpha\beta} \right\rangle_{\hat{w}})^2 \right\rangle_{\hat{w}} \tag{23}$$

と分散の形になる．以下，特に $\hat{w}$ として $\hat{w} = w^*$ という特別な場合について考える．特別な $\hat{w}$ を選ぶことで，平均場近似自由エネルギーの上限が得られる．

$\alpha \in \{K^* + 1, \cdots, K\}$ のとき，$w_{\alpha\beta}^* = 0$ で $t_{\alpha\beta} = 0$ となることから (23) 式は

$$\frac{\partial^2 H(w)}{\partial w_{\alpha\beta}^2}\Big|_{w^*} = 0 \quad for \ \alpha \in \{K^* + 1, \cdots, K\} \tag{24}$$

となる．したがって [補題] において $r = K^* M$ であり，全パラメータ数は $d = KM$ であることから，

$$\bar{F}(n) \leqq \frac{KM + K^* M}{4}\log n + C$$

が得られる．([定理] の証明終了)

5. 考　　察

5.1 事前分布について

「4. 主定理」では，事前分布を正規分布としたが，正規分布でない場合についてここで述べる．

(18) 式 $\varphi_0(\theta)$，関数 $\varphi(\theta)$ において $\varphi(\theta) \geqq c\varphi_0(\theta)$ を満たす定数 $c(> 0)$ が存在するとき，(7) 式 $\bar{F}(n)$ を事前分布の関数 $\bar{F}(\varphi)$ と表したもとで，$\log x$ の単調増加性をつかって，

$$\begin{aligned}\bar{F}(\varphi) &\leqq \bar{F}(c\varphi_0) = \bar{F}(\varphi_0) - \log c\\ &\leqq \frac{KM + K^* M}{4}\log n + C\end{aligned} \tag{25}$$

が成立する．したがって，$\varphi(\theta) \geqq c\varphi_0(\theta)$ を満たす定数 $c(> 0)$ が存在する $\varphi(\theta)$ においても「主定理 4.2」はそのまま成立する．

5.2 下界について

「4. 主定理」では，完全 2 部グラフ型ボルツマンマシンにおいて，平均場近似自由エネルギーの，学習サンプル数が増えたときの漸近形について上界を与えた．下界についてはまだ厳密な証明は得られていないが，ここでは試験分布 $\bar{f}(\theta)$ として正規分布族に限定した場合，さらにはその平均が真のパラメータ w^* に限定した場合について，下界を評価する．

(24) 式に対比して，二階微分係数は，$\alpha \in \{1, 2, \cdots, K^*\}$ のとき $w_{\alpha\beta}^* \neq 0$ であることから，

$$\frac{\partial^2 H(w)}{\partial w_{\alpha\beta}^2}\Big|_{w^*} \neq 0 \quad for \ \alpha \in \{1, 2, \cdots, K^*\} \tag{26}$$

である．これから二階微分係数の意味で，「4.2 主定理」のバウンドは最小である．

次に「補題」の結果である一般平均場近似自由エネルギーにおいてその上界を得るのに，(21) 式によって最適化パラメータ N_i を設定したが，(20) 式において，

$$\{ i ; \ \frac{\partial^2 H(\theta)}{\partial \theta_i^2}\Big|_{\hat{\theta}} = \frac{\partial^4 H(\theta)}{\partial \theta_i^2 \partial \theta_j^2}\Big|_{\hat{\theta}} = 0\} \quad for \ {}^{\forall} j \in \{1, \cdots, d\} \tag{27}$$

を満たす i が存在すれば，$N_i = n^{\frac{1}{3}}$ などと設定しても (20) 式は定数オーダーとできる．その結果，一般平均場近似自由エネルギーに対し，さらに小さい上界を与えることができる．そこで完全 2 部グラフ型ボルツマンマシンにおいて，$\hat{\theta} = w^*$ のとき，(27) 式を満たすパラメータ $N_{\alpha\beta}$ が存在するかを検討し，そのような $N_{\alpha\beta}$ は存在しないことを示す．それによって w^* を平均とする正規分布族の場合，「4.2 主定理」より小さくなる設定法は存在しないことを示す．(27) 式を満たすためには，(24) 式から，$\alpha \in \{K^* + 1, \cdots, K\}$ に限定できる．次に完全 2 部グラフ型ボルツマンマシンの四階微分は，任意の α,β について

$$\begin{aligned}\frac{\partial^4 H(w)}{\partial w_{\alpha\beta}^4} &= 8\{-\left\langle t_{\alpha\beta}^2 \right\rangle_{w^*} + \left\langle t_{\alpha\beta} \right\rangle_w^2\}\\ &\quad +6\{\left\langle t_{\alpha\beta}^4 \right\rangle_{w^*} - \left\langle t_{\alpha\beta} \right\rangle_w^4\}.\end{aligned} \tag{28}$$

$\delta \neq \beta$ のとき

$$\frac{\partial^4 H(w)}{\partial w_{\alpha\delta}^2 \partial w_{\alpha\beta}^2} = 2[\left\langle (3t_{\alpha\beta}^2 - 1)(t_{\alpha\beta}^2 - 1) \right\rangle_{w^*}$$

$$-\{3\left\langle t_{\alpha\beta}\right\rangle_w\left\langle t_{\alpha\delta}\right\rangle_w-\left\langle x_\beta x_\delta\right\rangle_w\}$$
$$\times\{\left\langle t_{\alpha\beta}\right\rangle_w\left\langle t_{\alpha\delta}\right\rangle_w-\left\langle x_\beta x_\delta\right\rangle_w\}]. \quad (29)$$

また $\gamma \neq \alpha$ のとき

$$\frac{\partial^4 H(w)}{\partial w_{\gamma\delta}^2 \partial w_{\alpha\beta}^2} = -2\{3\left\langle t_{\alpha\beta}\right\rangle_w\left\langle t_{\gamma\delta}\right\rangle_w-\left\langle t_{\alpha\beta}t_{\gamma\delta}\right\rangle_w\}$$
$$\times\{\left\langle t_{\alpha\beta}\right\rangle_w\left\langle t_{\gamma\delta}\right\rangle_w-\left\langle t_{\alpha\beta}t_{\gamma\delta}\right\rangle_w\} \quad (30)$$

である．$\hat{w} = w^*$ における微分係数は $\alpha \in \{K^*+1, \cdots, K\}$ のとき，$w^*_{\alpha\beta} = 0$ より $t_{\alpha\beta} = 0$ が成立することを考慮して (28), (30) 式は 0 となる．(29) 式は

$$\left.\frac{\partial^4 H(w)}{\partial w_{\alpha\delta}^2 \partial w_{\alpha\beta}^2}\right|_{w^*} = 2(1-\left\langle x_\beta x_\delta\right\rangle_{w^*}^2) \quad (31)$$

となる．ここで $\left\langle x_\beta x_\delta\right\rangle_{w^*}$ は，$\delta \neq \beta$ のとき

$$-\left\langle x_\beta x_\beta\right\rangle_{w^*} < \left\langle x_\beta x_\delta\right\rangle_{w^*} < \left\langle x_\beta x_\beta\right\rangle_{w^*}$$
$$\Longleftrightarrow \qquad -1 < \left\langle x_\beta x_\delta\right\rangle_{w^*} < 1 \quad (32)$$

より，(31) 式は 0 にならず，(27) を満たすパラメータ $N_{\alpha\beta}$ は存在しない．

5.3 統計的正則モデルとの比較

本論文では，完全 2 部グラフ型ボルツマンマシンにおいて，ベイズ事後分布を平易な正規分布に近似することによって平均場近似自由エネルギーの上界を与えた．今回ボルツマンマシンの全パラメータ数が KM であることから，統計的正則モデルの漸近論

$$F(n) = \frac{KM}{2}\log n + O(1)$$

と比較すれば，「4.2 主定理」は $K^* \leqq K$ から，小さい振舞いを持つことがわかる．

完全 2 部グラフ型ボルツマンマシンにおいては，平均場近似ではなくベイズ学習によるベイズ自由エネルギーの上界が，代数幾何学的な手法を用いて得られている [12]. [12] で導出された漸近形と本論文で導出した漸近形は一致している．このことは，ベイズ自由エネルギーの上界が，平易な正規分布による平均場近似で到達可能であることを意味する．

5.4 *SingIC*

特異モデルにおけるモデル選択として $SingIC$ が [13] で提案されている．$SingIC$ は，(3),(4) で与えられた λ, m が一般に学習モデルの隠れ素子数 K と真の隠れ素子数 K^* との関数であることから，

$$\lambda = F_1(K^*, K) \quad m = F_2(K^*, K) \quad (33)$$

とかけ，λ,m 依存の観測量 $y = g(\lambda, m)$ が得られたとき，

$$y = g(F_1(K^*, K), F_2(K^*, K)) \quad (34)$$

の方程式から，観測できない真の隠れ素子の個数 K^* を，推測する方法である．この枠組は，自由エネルギーの上界を用いても適用できることが提案され，関数 F_1, F_2 を数理的に導出することは，(34) を解く上で重要である．本論文の「主定理 4.1」は，正規分布族による自由エネルギーを利用した $SingIC$ においてその基礎につながるものである．

5.5 今後の課題

(1)4. 主定理では，$H(\hat{w}) = 0$ を満たす $\hat{w}$ として w^* のときだけを考えているが，$\hat{w} \neq w^*$ をすべて考えることで，試験分布として正規分布族に限定したときの平均場近似自由エネルギーの漸近形が厳密に求められる．

(2) 完全 2 部グラフ型ボルツマンマシンでなく，一般のボルツマンマシンに拡張して平均場近似自由エネルギーの漸近形を導出する．

(3) 正規分布族に限定した，平均場近似自由エネルギーを計算するアルゴリズムを作る．

6. むすび

完全 2 部グラフ型ボルツマンマシンの学習モデルにおいて，ベイズ事後分布を正規分布に近似することで，平均場近似自由エネルギーの上界を与えた．

文　献

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

[2] 青柳美輝, 渡辺澄夫, “縮小ランクモデルの汎化誤差と特異点解消,” 信学技報, Vol.104, No.7, pp.1029-1038, 2003.

[3] K. Yamazaki, and S. Watanabe, “Singularities in mixture models and upper bounds of stochastic complexitiy,” International Journal of Neural Networks, 16(2003), pp.1029-1038, 2003.

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

[5] S. Nakajima, and S. Watanabe, “Generalization Error of Variational Bayes Approach in Reduced Rank Regression,” 信学技報, Vol.104, No.760, pp.117-122, 2004.

[6] 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.

[7] 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, Vol.104, No.760, pp189-194, 2004.

[8] 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.

[9] 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, Vol.104, No.760, pp.111-116, 2005.

[10] S. Amari, and N. Murata, “Statistical theory of learning curves under entropic loss,” Neural Computation, Vol.5, pp.140-153, 1993.

[11] K. Yamanishi, “A decision-theoretic extension of stochastic complexity and its applications to learning,” IEEE Trans. Information Theory, Vol.44, No.4, pp.1424-1439, 1998.

[12] K. Yamazaki, and S. Watanabe, “Singularities in complete bipartite graph-type boltzmann machines and upper bounds of stochastic complexities,” IEEE Trans. Neural Networks, Vol.16(2), pp.312-324, 2003.

[13] 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, Vol.105, No.211, pp.7-12, 2005.