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

完全２部グラフ型ボルツマンマシンのベイズ予測誤差

Bayes Generalization Errors of Complete Bipartite Graph-type Boltzmann Machines

山崎 啓介* 渡辺 澄夫†
Keisuke Yamazaki Sumio Watanabe

Abstract: It is well known that Boltzmann machines are non-regular statistical models. The set of their parameters of small size models is an analystic set with singularities in the space of a large size ones. The mathematical foundation of their learning is not yet constructed because of the singularities, though they are applied in many situations of information engineering. Recently we established the method to calculate the bayes generalization errors with an algebraic geometric method even if the models are non-regular. This paper shows that the upper bounds of generalization errors in Boltzmann machines is smaller than the ones in regular statistical models.

1 はじめに

ニューラルネットワーク, ボルツマンマシン, 混合正規分布などの統計的推論モデルはパターン認識, システム制御, 時系列予測など情報工学の分野で多くの応用例をもっている. 統計的正則モデルと呼ばれるモデルは統計学で長く研究されてきた. その数理的な性質は 1970 年代にほぼ解明され, モデル選択規準や予測の精度はよく知られている. しかしながら, 情報工学で用いられるモデルはそのほとんどが特定不能なモデルであり正則モデルではないため, 数理的な基盤が確立されていない.

1 つのパラメータを空間上の 1 点とみなすと, モデルが取りうるパラメータ全体はパラメータ空間として表現することができる. 特定不能なモデルの場合, このパラメータ空間において, そのモデルよりも規模の小さなモデルを表現するパラメータは広がりをもつ解析的集合として表現される. この集合上の点ではフィッシャー情報行列が縮退し, 一般的に特異点をもつ. このため特異モデルとも呼ばれる. フィッシャー情報行列はモデルの解明に重要な役割を果たす量であり, 統計的な手法ではその逆行列を用いて予測精度などが表現される. このため特異モデルでは統計的手法が適用できない. 近年では, 最ゆう推定量の漸近的振る舞いについて情報幾何 [2], 経験確率過程論, 順序統計量, 多変数関数論などの様々なアプローチにより, 特異モデルが統計的正則モデルとは異なる特徴をもつことが解明されつつある.

そのようなアプローチとは別に, ベイズ予測に関しては, 特異モデルであってもモデルの予測精度が解明できることが示された [4], [5]. 真の分布が学習モデルに含まれるとき, 汎化誤差は代数幾何学における特異点論と密接な関係を持つことが知られている.

そこで本論文ではグラフィカルモデルの一部であるボルツマンマシンにおいて, 最も基礎的で解析が容易と思われる完全２部グラフの結合を持つものを考える. このモデルに代数幾何的な手法を適用することで, その予測精度を明らかにする. ベイズ予測において, 特異モデルは同じ規模の統計的推測モデルに比べ優れた予測精度をもつことは知られているが, 本論文は具体的なモデルで実際の計算を行うことにより, それを定量的に示すものである.

2 学習理論

ここでは, ベイズ予測と学習曲線について, 既に知られていることをまとめる.

*東京工業大学大学院総合理工学研究科, 〒 226-8503 横浜市緑区長津田 4259, e-mail zaki23@pi.titech.ac.jp,
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, 4259 Nagatsuta, Midori-ku, Yokohama, 226-8503 Japan
†東京工業大学 精密工学研究所, swatanabe@pi.titech.ac.jp,
Precidion and Intelligence Laboratory, Tokyo Institute of Technology,

2.1 ベイズ予測

入出力 x の空間を R^M とし, パラメータ w の空間を $W \subset R^d$ とする. 学習モデルは入出力に対する確率密度関数であり, その性質はモデルのもつパラメータにより決定されるので $p(x|w)$ と表記する. 真の分布 $q(x)$ に従う n 個のデータサンプル $X^n = (X_1, X_2, \cdots, X_n)$ を用いてモデルの学習を行う. 本論文では学習モデルは真の分布を含むとする. つまり真のパラメータ w^* が存在し $p(x|w^*) = q(w)$ が実現できる. 特異モデルでは w^* は 1 点にはならず解析的集合となる.

ベイズ予測はパラメータについての事前分布 $\varphi(w)$ を定め, サンプル X^n と事前分布を用いて, 事後確率

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

を定める. ただし Z_n は正規化定数

$$Z_n = \int \varphi(w)\prod_{i=1}^{n} p(X_i|w)dw$$

である. この事後確率によって新たな入出力 x に対する予測を

$$p(x|X^n) = \int p(x|w)p(w|X^n)dw$$

で行うのがベイズ予測である. $p(x|X^n)$ はベイズ予測分布と呼ばれる.

2.2 特異モデルの学習理論

真の分布から学習モデルまでの距離概念として, カルバック情報量 $H(w)$ と経験カルバック情報量 $H_n(w)$ をそれぞれ

$$\begin{aligned} H(w) &= \int q(x)\log\frac{q(x)}{p(x|w)}dx, \\ H_n(w) &= \frac{1}{n}\sum_{i=1}^{n}\log\frac{q(X_i)}{p(X_i|w)} \end{aligned}$$

と定義する. 真の分布からベイズ予測分布までのカルバック情報量は

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

となり汎化誤差と呼ばれる. $E_n\{\cdot\}$ はサンプル X^n の現れ方について平均をとることを意味する. 統計的正則モデルでは $G(n)$ はモデルに依存せず, モデルのパラメータ数を d として $G(n) = d/2n$ となることが知られている. しかしながら, 情報工学で用いられる神経回路網や混合分布モデル, ボルツマンマシンなどでは異なる.

確率的複雑さは

$$F(n) = -E_n\left\{\log\int \exp(-nH_n(w))\varphi(w)dw\right\} \quad (1)$$

で定義され, 汎化誤差との関係は

$$G(n) = F(n+1) - F(n)$$

となることが知られている [1],[8]. つまり予測精度を計算するという問題は, 確率的複雑さを計算することに帰着する.

$H(w)$ が実解析関数で $\varphi(w)$ がコンパクトサポートであるとする. このとき $Re(z) > 0$ で正則な 1 変数複素関数を

$$J(z) = \int H(w)^z\varphi(w)dw$$

と定義すると, この関数は複素平面全体に有理型関数として解析接続が可能であり, その極は実軸の負の領域にのみ存在するという性質をもつ. 本論文では $J(z)$ を $H(w), \varphi(w)$ により定まるゼータ関数と呼ぶ. $J(z)$ の最も原点に近い極とその位数をそれぞれ $-\lambda, m$ とおくと, 確率的複雑さは

$$F(n) = \lambda\log n - (m-1)\log\log n$$

という漸近展開をもつ [4]. よって汎化誤差は漸近展開をもつとすると,

$$G(n) = \frac{\lambda}{n} - \frac{m-1}{n\log n}$$

となる.

λ, m は解析的集合 $\{w; H(w) = 0\}$ に関して, 特異点解消という手法を用いることにより, 計算することができる [5],[9]. 高度で大規模な学習モデルにおいて完全な特異点解消を見出すことは困難を伴うことが多いが, 部分的な特異点解消によって λ の上限を得ることができる. 本論ではこれまでに理論的解明がなされていない完全 2 部グラフ結合を持つボルツマンマシンに関して, 部分的な特異点解消 (ブローアップ) を適用することで, 汎化誤差の主要項の係数である λ を明らかにする.

3 結果

ここでは **2** で述べた方法を, 混合分布モデルとボルツマンマシンに適用し, それぞれのモデルにおけるベイズ汎化誤差の上限を与える定理を述べ, その証明を行う.

3.1 準備

ゼータ関数とその極について次のような補題が知られている.

補題 1 $\varphi_1(w)$,$\varphi_2(w)$ により定まる $H(w)$ のゼータ関数

$$J_i(z) = \int H(w)^z \varphi_i(w) dw \qquad (i = 1, 2)$$

の最も原点に近い極をそれぞれ $-\lambda_1$,$-\lambda_2$ とする．このとき,

$$\varphi_1(w) \geq \varphi_2(w) \text{ ならば} \lambda_1 \leq \lambda_2.$$

補題 2 $H(w)$,$H_1(w_1)$,$H_2(w_1)$ より定まるゼータ関数

$$\begin{aligned} J(z) &= \int H(w)^z \varphi(w) dw, \\ J_i(z) &= \int H_i(w_i)^z \varphi(w_i) dw_i \qquad (i = 1, 2) \end{aligned}$$

の最も原点に近い極をそれぞれ $-\lambda$,$-\lambda_1$,$-\lambda_2$ とする．$w = \{w_1, w_2\}$ であれば一般に

$$H(w) \leq H_1(w_1) + H_2(w_2) \ \ (\forall w) \text{ ならば } \lambda \leq \lambda_1 + \lambda_2.$$

補題 1 における $\varphi_2(w_2)$ は分布関数でなくとも成立する．これは J_1 の積分範囲を限って J_2 とした場合,λ_2 は λ_1 の上限となることを意味する．

以上の 2 つの補題を用いて定理の証明を行う．

3.2 ボルツマンマシン

ここでは本論文が扱うボルツマンマシンを定式化し, そのベイズ汎化誤差の上限を与える定理を述べ, その証明を行う．

3.2.1 ボルツマンマシンの汎化誤差

入出力素子を $x = \{x_j\}_{j=1}^M \in \{-1, +1\}^M$ とし, 隠れ素子を $h = \{h_i\}_{i=1}^K \in \{-1, +1\}^K$ とする．本研究では入出力素子間や隠れ素子間の結合は存在せず, 入出力素子と隠れ素子の間が全結合されている形状を扱うことにする. つまり各入出力 x_j は全ての隠れ素子と結合している. グラフ理論ではこの形状を完全 2 部グラフと呼ぶ. モデルのパラメータは $w = \{u_{ij}\} \in R^{K \times M}$ であり u_{ij} で j 番目の入出力素子から i 番目の隠れ素子への結合重みを表すものとする. この形状のボルツマンマシン $p(x|w)$ は

$$p(x|w) = \frac{1}{Z(w)} \prod_{i=1}^{K} (e^{-\sum_{j=1}^M u_{ij} x_j} + e^{\sum_{j=1}^M u_{ij} x_j})$$

と表される. 簡単のため $\sum_{x_1=\pm 1} \cdots \sum_{x_M=\pm 1}$ を $\sum_x$ と略記すると,

$$Z(w) = \sum_x \prod_{i=1}^{K} (e^{-\sum_{j=1}^M u_{ij} x_j} + e^{\sum_{j=1}^M u_{ij} x_j})$$

で正規化されている. この $p(x|w)$ を用いて, 真の分布

$$q(x) = \frac{1}{Z^*(w^*)} \prod_{i=1}^{H} (e^{-\sum_{j=1}^M u_{ij}^* x_j} + e^{\sum_{j=1}^M u_{ij}^* x_j})$$

を学習することを考える. ただし,

$$Z^*(w^*) = \sum_x \prod_{i=1}^{H} (e^{-\sum_{j=1}^M u_{ij}^* x_j} + e^{\sum_{j=1}^M u_{ij}^* x_j})$$

であり $w^* = \{u_{ij}^*\}$ は真のパラメータで定数とし $H < K$ とする. このとき次の定理が成立する．

定理 1 ボルツマンマシンの汎化誤差 $G(n)$ は次の上限を持つ．

$$\begin{aligned} G(n) &\leq \frac{\lambda}{n} \\ \lambda &= \begin{cases} (K + 3H + 1)/4 & (\ if\ M = 2) \\ (K + H)M/4 & (M \geq 3) \end{cases} \end{aligned}$$

3.2.2 定理 1 の証明

まず, 次の補題が成立する．

補題 3 学習モデルと真の分布がそれぞれ $\rho(x, w) > 0$ を用いて

$$\begin{aligned} p(x|w) &= \frac{\rho(x, w)}{Z(w)}, \\ q(x) &= p(x|w^*) = \frac{\rho(x, w^*)}{Z(w^*)} \end{aligned}$$

で表されているとする．$\rho(x, w)$ については 2 つの積にパラメータごと分離でき, $w = \{w_1, w_2\}$ において

$$\rho(x, w) = \rho_1(x, w_1) \rho_2(x, w_2)$$

とする. 正規化はそれぞれ

$$\begin{aligned} Z(w) &= \sum_x \rho(x, w), \\ Z_i(w_i) &= \sum_x \rho_i(x, w_i) \qquad (i = 1, 2) \end{aligned}$$

のように表記する. このとき $w = w^*$ の近傍において, 十分大きな定数 L を用いて

$$H(w) \leq L\,[H_1(w_1) + H_2(w_2)]$$

のように, カルバック情報量を分離した形で評価できる．ここで,

$$\begin{aligned} H(w) &= \sum_x \frac{\rho(x, w^*)}{Z(w^*)} \log \frac{Z(w) \rho(x, w^*)}{Z(w^*) \rho(x, w)}, \\ H_i(w_i) &= \sum_x \frac{\rho_i(x, w_i^*)}{Z_i(w_i^*)} \log \frac{Z_i(w_i) \rho_i(x, w_i^*)}{Z_i(w_i^*) \rho_i(x, w_i)} \quad (i = 1, 2) \end{aligned}$$

とし,$w = w^*$ の近傍では $w_1 \to w_1^*, w_2 \to w_2^*$ と仮定した．

(補題 3 の証明は紙面の都合上省略する)

ボルツマンマシンは補題 3 の仮定を満たすので, 補題 3 を適用することにより,

$$
\begin{aligned}
H(w) &\leq L\left[\sum_x \frac{\rho_1(x,w_1^*)}{Z_1(w_1^*)}\log\frac{Z_1(w_1)\rho_1(x,w_1^*)}{Z_1(w_1^*)\rho_1(x,w_1)}\right. \\
&\quad \left.+\sum_x \frac{\rho_2(x,w_2^*)}{Z_2(w_2^*)}\log\frac{Z_2(w_2)\rho_2(x,w_2^*)}{Z_2(w_2^*)\rho_2(x,w_2)}\right], \quad (2) \\
Z_i(w_i) &= \sum_x \rho_i(x,w_i)
\end{aligned}
$$

が成立する. ここで

$$
\begin{aligned}
\rho(x,w) &= \prod_{i=1}^{K}(e^{-\sum_{j=1}^{M} u_{ij}x_j} + e^{\sum_{j=1}^{M} u_{ij}x_j}), \\
\rho_1(x,w_1) &= \prod_{i=1}^{H}(e^{-\sum_{j=1}^{M} u_{ij}x_j} + e^{\sum_{j=1}^{M} u_{ij}x_j}), \\
\rho_2(x,w_2) &= \prod_{i=H+1}^{K}(e^{-\sum_{j=1}^{M} u_{ij}x_j} + e^{\sum_{j=1}^{M} u_{ij}x_j})
\end{aligned}
$$

とおいた. つまり $w_1 = \{u_{ij} : 1 \leq i \leq H, 1 \leq j \leq M\}$,$w_2 = \{u_{ij} : H+1 \leq i \leq K, 1 \leq j \leq M\}$ とみなした. (2) の第 1 項目を $H_1(w_1)$, 第 2 項目を $H_2(w_2)$ とおくと $H_1(w_1)$ は真の分布と同じ数の隠れ素子を持つモデルで学習したときのカルバック情報量を意味しており $H_2(w_2)$ は 0 を真のパラメータとする分布を学習したときのカルバック情報量となっている. そこで補題 2 よりそれぞれのカルバック情報量から定まるゼータ関数の極を求めればよい. まず $H_1(w_1)$ においては次の補題が成り立つ.

補題 4 $H_1(w_1)$ が表現するモデルに対する汎化誤差を $G_1(n)$ とおくと,

$$
\begin{aligned}
G_1(n) &\leq \frac{\lambda_1}{n} \\
\lambda_1 &= \frac{HM}{2}
\end{aligned}
$$

が成り立つ.

[補題 4 の証明] $H_1(w_1)$ は真の分布と学習モデルの隠れ素子の数が同じである. よって $q(x) = p(x|w)$ となるパラメータは $w = w^*$ のみとなり特定可能であるので, 統計的正則モデルのカルバック情報量を表している. パラメータの数が HM より補題 4 が証明された. (補題 4 の証明終)

次に $H_2(w_2)$ が定めるゼータ関数の極を求める. 以下では真のパラメータが $w^* = 0$ の場合を考える. つまり $q(x)$ は一様分布とする. まず入出力の次元 $M = 2$ のとき, 次の補題が成り立つ.

補題 5 $M = 2$ のときの $H_2(w_2)$ が表現するモデルの汎化誤差を $G_2(n)$ とおくと,

$$
\begin{aligned}
G_2(n) &\leq \frac{\lambda_2}{n} \\
\lambda_2 &= \frac{K-H+1}{4}
\end{aligned}
$$

が成り立つ.

[補題 5 の証明] 真の分布は $M = 2$ より

$$q(x) = \frac{1}{4}$$

となる. よってカルバック情報量は cosh を用いて表記でき, さらに cosh の加法定理

$$\cosh(\alpha \pm \beta) = \cosh(\alpha)\cosh(\beta) \pm \sinh(\alpha)\sinh(\beta)$$

を用いて展開した後 x について和をとることで

$$
\begin{aligned}
H_2(w_2) &= -\frac{1}{2}\log\left[1 - T(w_2)^2\right], \\
T(w_2) &= \left(\frac{S_{odd}(w_2)}{\prod_{i=H+1}^{K}\cosh u_{i1}\cosh u_{i2} + S_{even}(w_2)}\right)^2
\end{aligned}
$$

が成立する. ただし $\prod_i \cosh(u_{i1}+u_{i2})$ を加法定理で展開した項の中で, 因数 $\sinh u_{i1}\sinh u_{i2}$ を奇数個, または偶数個含む項の和をそれぞれ $S_{odd}(w_2), S_{even}(w_2)$ とおいた. 補題 1 より $w = 0$ の近傍を考えれば十分なので, 十分小さな $\epsilon > 0$ をとって, パラメータの集合 $W_2' = \|u_{ij}\| \leq \epsilon$ 内を考える. この近傍では $S_{even}(w_2) \to 0$,$S_{odd}(w_2) \to 0$ より, $\log(1+x) \simeq x \ \ (x \to 0)$ から

$$H_2(w_2) \simeq \frac{1}{2}\left(\frac{S_{odd}(w_2)}{\prod_{i=H+1}^{K}\cosh u_{i1}\cosh u_{i2}}\right)^2$$

となる. また $\tanh x \simeq x \ \ (x \to 0)$ より

$$H_2(w_2) \simeq \frac{1}{2}\left(T_{odd}(w_2)\right)^2$$

が成り立つ. ただし $T_{odd}(w_2)$ は w_2 の中で奇数個の $u_{i1}u_{i2}$ を因数にもつ項の和である.W_2' の内部では定数 c_1, c_2 が存在して

$$\frac{1}{2}\left(T_{odd}(w_2)\right)^2 \leq c_1\left(\sum_{i=H+1}^{K} u_{i1}u_{i2}\right)^2 + c_2 T_{Comb}(w_2)$$

が成立するようにとれる. ただし $T_{Comb}(w_2)$ は $u_{H+1,1}$ を除く u_{i1} の中から 2 つ選んだ全ての組合せの 2 乗和である.

$$H_3(w_2) = c_1\left(\sum_{i=H+1}^{K} u_{i1}u_{i2}\right)^2 + c_2 T_{Comb}(w_2)$$

とおくと $H_2(w_2) \leq H_3(w_2)$ なので, 補題 1 と補題 2 より

$$J_1 = \int_{W_2'} H_3(w_2)^z \varphi(w_2) dw_2$$

の極の絶対値の上限を求めればよい.

$$g : w_3 = (\nu, \{\nu_i\}_{i=H+2}^K, \{\mu_i\}_{i=H+1}^K) \mapsto w_2$$

と写像を定義し

$$\begin{aligned} u_{11} &= \nu^2 - \sum_{i=H+2}^K \nu_i \mu_i, \\ u_{i1} &= \nu_i \nu \ \ (i = H+2, H+3, \cdots, K), \\ u_{H+1,2} &= \mu_{H+1}, \\ u_{i2} &= \mu_i \mu_1 \ \ (i = H+2, H+3, \cdots, K) \end{aligned}$$

とおくと, ヤコビアンは

$$|g'| = \nu^{K-H}$$

となるので, 上限として

$$\lambda_2 = \frac{K-H+1}{4}$$

が得られた. (補題 5 の証明終)

次に入出力次元 $M \geq 3$ の場合を考える. 隠れ素子が 1 つの場合には, 次の補題が成り立つ.

補題 6 $M \geq 3$,$K - H = 1$ のときの $H_2(w_2)$ が表現するモデルの汎化誤差を $G_2(n)$ とおくと,

$$\begin{aligned} G_2(n) &\leq \frac{\lambda_2}{n} \\ \lambda_2 &= \frac{M}{4} \end{aligned}$$

が成り立つ.

[補題 6 の証明] $w_2^* = 0$ より真の分布が

$$q(x) = \frac{1}{2^M}$$

の場合を考える. よってカルバック情報量は $u_{H+1,j}$ を u_j と略記することで

$$H_2(w_2) = -\frac{1}{2^M} \sum_x \log \frac{2^M \cosh\left(\sum_{j=1}^M u_j x_j\right)}{\sum_x \cosh\left(\sum_{j=1}^M u_j x_j\right)}$$

と表記できる. 分母において x について和をとると, 加法定理より

$$H_2(w_2) = -\frac{1}{2^M} \sum_x \log \frac{\cosh\left(\sum_{j=1}^M u_j x_j\right)}{\prod_{j=1}^M \cosh(u_j)}$$

となる. 分子においても x について和をとることを考える. これは

$$f(w_2) = \prod_x \cosh\left(\sum_{j=1}^M u_j x_j\right)$$

を求めることに帰着される. 加法定理を用いて展開し, 実際に x を代入して計算すると, 定数 c_1 により

$$H_2(w_2) \leq -\frac{1}{2^M} \log\left(1 - c_1 \sum_{k,l}^M (\tanh u_k \tanh u_l)^2\right)$$

となる. $x \to 0$ において $\log(1+x) \simeq x$ と $\tanh x \simeq x$ より, 定数 c_2 を用いて

$$H_2(w_2) \leq H_3(w_2) = c_2 \sum_{k,l}^M (u_l u_k)^2$$

が成立する. 補題 1 と補題 2 より

$$J_2 = \int_{W_2'} H_3(w_2)^z \varphi(w_2) dw_2$$

の極の絶対値の上限を求めればよい.

$$g : w_3 = (\nu, \{\nu_j\}_{j=2}^M) \mapsto w_2$$

と写像を定義し,

$$\begin{aligned} u_1 &= \nu, \\ u_j &= \nu_j \nu \ \ (j = 2, 3, \cdots, M) \end{aligned}$$

とおいた. ヤコビアンは

$$|g'| = \nu^{M-1}$$

となるので, 上限として

$$\lambda_2 = \frac{M}{4}$$

が得られた. (補題 6 の証明終)

補題 6 より, 補題 2,3 を用いて隠れ素子が $K-H$ 個の場合のゼータ関数の極の絶対値の上限は $M(K-H)/4$ であることがわかる. 補題 4,5,6 より, 補題 2 を適用することで,

$$\begin{aligned} \lambda &\leq \lambda_1 + \lambda_2 \\ &= \begin{cases} (K+3H+1)/4 & (\text{ if } M = 2) \\ (K+H)M/4 & (M \geq 3) \end{cases} \end{aligned}$$

が得られた. (定理 1 の証明終)

4 考察

本論文でのボルツマンマシンは完全２部グラフの結合をもつものである. 本来のボルツマンマシンは全結合のモデルであり, 本論文の方法をそのまま拡張することはできないが, 全結合ボルツマンマシンの解明の基礎となるモデルである. 特に入出力素子間に結合がある場合は, 本論文と全く同様の議論により次の系が成り立つ.

系 1 入出力次元が M, 隠れ素子と入出力素子の間が全結合しており, さらに入出力素子間も全結合しているボルツマンマシンにおいて, 真の分布を隠れ素子数 H 個とし, これを隠れ素子数 K 個のモデルで学習したとき, 汎化誤差は

$$
\begin{aligned}
G(n) &\leq \frac{\lambda}{n} \\
\lambda &= \begin{cases} (K+3H+3)/4 & (\ if\, M=2) \\ (K+H+M-1)M/4 & (M \geq 3) \end{cases}
\end{aligned}
$$

となる.

補題 3 より入出力素子間の全結合と, 完全２部グラフの結合にモデルを分けて考えることができる. 入出力素子の全結合は正則モデルとみなすことができるので, そのパラメータ数は $M(M-1)/2$ となり, この部分の汎化誤差係数は $M(M-1)/4$ であることがわかる. よって定理１の結果にこれを加えればよい.

真の分布に対し学習モデルは隠れ素子が多い. これは冗長な部分が生じることを意味する. 本論文の結果はこの冗長な部分の汎化誤差係数が, パラメータ数を d とすると $d/4$ となり $d/2$ の統計的正則モデルに比べ, 汎化誤差の係数は小さいことが明らかになった. このことはモデル選択と汎化誤差が異なる問題であることを示している. 本論文の結果によりモデル選択の規準として広く知られている Schwarz の BIC が, モデル選択に一般的には適用できないことがわかる. ただし, 本論文では特異点の周りでコンパクトサポートになるような事前分布を考えている. 真のパラメータ上で 0 となるような Jeffreys' prior を用いれば BIC が適用できることが知られている.[7]. Jeffreys' prior を使って予測を行う場合, $\lambda = d/2$ となることが知られており, 予測精度の意味で不向きではあるがモデル選択には有効である.

また, 真の分布が含まれるという仮定から計算される本論文の結果は, その仮定が成り立たないような場合の基礎となるものである [6].

5 まとめ

完全２部グラフ結合をもつボルツマンマシンのベイズ汎化誤差の上限を定理として証明した. これにより統計的正則モデルとの定量的な比較が可能となり, モデルの汎化誤差は正則モデルと比べ, はるかに小さいことを確認した.

謝辞

本研究の一部は文部省科学研究費補助金 12680370 によって行われた.

参考文献

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

[2] S. Amari and T. Ozeki, "Differential and algebraic geometry of multilayer perceptrons," *IEICE Trans.*, to appear.

[3] Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461-464.

[4] S. Watanabe, "Algebraic analysis for non-identifiable learning machines," *Neural Computation*, Vol.13, No.4, pp.899-933, 2001.

[5] 渡辺澄夫, "特異点を持つ学習モデルと事前分布の代数幾何," 人工知能学会誌、Vol.16, No.2, pp.308-315, 2001.

[6] S. Watanabe, "Algebraic geometrical methods for hiearchical learning machines," Neural Networks, Vol.14, No.8,pp.1049-1060, 2001.

[7] S. Watanabe, "Algebraic information geometry for learning machines with singularities," Advances in Neural Informaion Processing Systems, MIT Press, 14,pp.329-336, 2001.

[8] K. Yamanishi, "A decision-theoretic extension of stochastic complexity and its applications to learning," *IEEE Transactions on Information Theory*, Vol. 44, No.4, pp.1424-1439, 1998.

[9] 山崎啓介, 渡辺澄夫, "特異点を持つ推論モデルの確率的学習精度計算アルゴリズム," 信学技報 NC2000-64, 23-30, 2000.

[10] 山崎啓介, 渡辺澄夫, "特異点解消による混合正規分布のベイズ汎化誤差の解明," 日本神経回路学会講演論文集, pp.11-12, 2001.

[11] 山崎啓介, 渡辺澄夫, "特異点を持つ推論モデルの学習曲線の確率的計算法," 電子情報通信学会論文誌, J85-D-II, No.3, pp.363-372, pp.11-12, 2002.

[12] 山崎啓介, 渡辺澄夫, "混合分布モデルにおける確率的複雑さの解明," 信学技報 NC2001-150, 135-142, 2002.

[13] 山崎啓介, "特異点解消によるグラフィカルモデルのベイズ汎化誤差の解明," 情報論的学習理論 (IBIS2002) 予稿集, to appear.