2004 年情報論的学習理論ワークショップ
2004 Workshop on Information-Based Induction Sciences (IBIS2004)
Tokyo, Japan, November 8 - 10, 2004.

経験ベイズ法の汎化誤差について
Generalization Error of an Empirical Bayes Approach

中島 伸一* 渡辺 澄夫†
Shinichi Nakajima Sumio Watanabe

Abstract: To clarify the relationship between an empirical Bayes approach and the Bayes estimation, we analyzed the generalization error of an empirical Bayes approach in the case that the learning machine is layered and has singularities. As a result, we have found that behavior of the generalization error of the empirical Bayes approach and that of the Bayes estimation are similar to each other. We discuss the similarity and the difference between them.

Keywords: Empirical Bayes, Layered model, Stein estimator, Generalization error

1 はじめに

神経回路網，混合正規分布，隠れマルコフモデル，縮小ランク回帰などのモデルは，パターン認識，時系列予測，システム制御などに広く応用されている．これらのモデルは Fisher 計量が縮退する特異点を持つため，統計的正則モデルの漸近論を適用することができず，解析が困難であったが，近年，代数幾何学を用いる解析法が確立された [1–3]．これを期に，多くの特異モデルにおいて，真の分布が特異点に含まれる場合の学習係数の厳密な値，あるいはその上限値がこの方法によって解明されてきた [4–9]．これらのすべての解析結果は，特異点の存在が，モデルが冗長であることによる過学習を抑制し，正則モデルと比較してベイズ汎化誤差が著しく小さくなることを示している．

真の分布が厳密に特異点上にない場合の解析も行われている [10]．そこでは，真の分布と特異点とのカルバック擬距離がサンプル数に反比例するスケーリングを用いることにより，真の分布が特異点に含まれる場合（この場合は上記の代数幾何学的解析手法が適用できる）と，真の分布が特異点から十分離れた場合（この場合はモデルに冗長な部分がなくなり，学習係数は正則モデルのそれと同じになる）との間を連続的に解析している．その結果，パラメータの次元が大きくなると，真の分布がどこにあっても常に正則モデルより優れた推定ができるという，Stein 推定量 [13, 14] に似た性質が見出された．

一方，学習モデルにハイパーパラメータを導入し，その値を周辺尤度によって決定する方法は，経験ベイズ法と呼ばれるが，Stein 推定量はこの経験ベイズ法による推定量として解釈されることが知られている [11, 12]

本論文では，経験ベイズ法とベイズ推定との関係をより明確にするために，経験ベイズ法の汎化誤差を [10] と同様のスケーリングを用いることによって解明し，ベイズ法との類似性，相違点について議論する．

まずはじめに 2 節において問題設定を行い，3 節で経験ベイズ法の枠組みについて解説する．4 節および 5 節ではそれぞれ，汎化誤差および学習誤差の解析を行う．6 節では経験ベイズ法とベイズ推定の類似性と相違点について考察し，工学的応用の可能性について議論する．最後に 7 節でまとめと今後の課題について述べる．

2 問題設定

K 次元の確率ベクトル変数 $\boldsymbol{x} \in R^K$ が，平均ベクトル $\boldsymbol{w}^* \in R^K$，共分散行列 $\mathbf{1}^K$（$K \times K$ 単位行列）の K 次元正規分布に従うとする．すなわち $\boldsymbol{x}$ の真の分布は

$$q(\boldsymbol{x}) = \frac{1}{(2\pi)^{K/2}} \exp\left(-\frac{|\boldsymbol{x} - \boldsymbol{w}^*|^2}{2}\right) \qquad (1)$$

である．これを $a \in R$，$\boldsymbol{b} \in R^K$ をパラメータとして持つ，以下のような階層モデルで学習することを考える．

$$p(\boldsymbol{x}|a, \boldsymbol{b}) = \frac{1}{(2\pi)^{K/2}} \exp\left(-\frac{|\boldsymbol{x} - a\boldsymbol{b}|^2}{2}\right) \qquad (2)$$

*東京工業大学総合理工学研究科, 〒 226-8503 神奈川県横浜市緑区長津田 4259
Tokyo Institute of Technology, 4259 Nagatsuda, Midori-ku, Yokohama-shi, Kanagawa, 226-8503 Japan
（株）ニコン, 〒 360-8559 埼玉県熊谷市大字御稜威ヶ原 201-9
Nikon Corporation, 201-9 Miizugahara, Kumagaya-shi, Saitama, 360-8559 Japan
tel. 045-924-5018, e-mail nakajima.s@cs.pi.titech.ac.jp
†東京工業大学総合理工学研究科, Tokyo Institute of Technology, swatanab@pi.titech.ac.jp

このモデルは，$\boldsymbol{w} = a\boldsymbol{b}$ と変数変換すれば正則モデルであるので，最尤推定法について議論する立場からは特異モデルには分類されないが [16]，ベイズ推定においては他の特異モデルと同様に，冗長性による過学習を特異点が抑制する効果を示す．

ここで，モデル (2) のパラメータの一部をハイパーパラメータとみなすことにより，経験ベイズ法が適用できる．本論では，a をハイパーパラメータ，$\boldsymbol{b}$ を（周辺化されるべき）パラメータとみなす場合を考える．以下，このことを明示するため，ハイパーパラメータは ; の後に書く．すなわち学習モデルの確率密度は

$$p(\boldsymbol{x}|\boldsymbol{b};a) = \frac{1}{(2\pi)^{K/2}} \exp\left(-\frac{|\boldsymbol{x}-a\boldsymbol{b}|^2}{2}\right) \tag{3}$$

である．$\boldsymbol{b}$ の事前分布を

$$\phi(\boldsymbol{b}) = \frac{1}{(2\pi)^{K/2}} \exp\left(-\frac{|\boldsymbol{b}|^2}{2}\right) \tag{4}$$

とする[1]．

3 経験ベイズ法

3.1 ベイズ推定

n 個のサンプル $\boldsymbol{X}^n = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ が得られた時，モデル (3)，(4) の周辺尤度は，平均ベクトル $\boldsymbol{m}(\boldsymbol{X}^n) = \sum \boldsymbol{x}_i/n$ を用いて

$$\begin{aligned} Z(\boldsymbol{X}^n;a) &= \int \left(\prod_{i=1}^n p(\boldsymbol{x}_i|\boldsymbol{b};a)\right)\phi(\boldsymbol{b})d\boldsymbol{b} \\ &= \frac{\int \exp\left(-\frac{1}{2}\sum_{i=1}^n |\boldsymbol{x}_i - a\boldsymbol{b}|^2 - \frac{|\boldsymbol{b}|^2}{2}\right)d\boldsymbol{b}}{(2\pi)^{(n+1)K/2}} \\ &= \frac{\exp\left(-\frac{1}{2}\sum|\boldsymbol{x}_i|^2\right)}{(2\pi)^{nK/2}} \cdot \frac{\exp\left(\frac{n^2a^2|\boldsymbol{m}|^2}{2(na^2+1)}\right)}{(na^2+1)^{K/2}} \end{aligned} \tag{5}$$

と表される．事後分布は

$$\begin{aligned} p(\boldsymbol{b}|\boldsymbol{X}^n;a) &= \frac{\left(\prod_{i=1}^n p(\boldsymbol{x}_i|\boldsymbol{b};a)\right)\phi(\boldsymbol{b})}{Z(\boldsymbol{X}^n;a)} \\ &= \left(\frac{na^2+1}{2\pi}\right)^{K/2} \exp\left(-\frac{1}{2}\left(na^2+1\right)\left|\boldsymbol{b} - \frac{na\boldsymbol{m}}{na^2+1}\right|^2\right) \end{aligned}$$

であり，予測分布は

$$\hat{\boldsymbol{w}}(a) = \frac{na^2\boldsymbol{m}}{na^2+1} \tag{6}$$

$$\hat{\sigma}^2(a) = \frac{na^2+a^2+1}{na^2+1} = 1 + \frac{1}{na^2+1} \tag{7}$$

を用いて

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{X}^n;a) &= \int p(\boldsymbol{x}|\boldsymbol{b};a)p(\boldsymbol{b}|\boldsymbol{X}^n;a)d\boldsymbol{b} \\ &= \frac{1}{(2\pi\hat{\sigma}^2)^{K/2}} \exp\left(-\frac{1}{2\hat{\sigma}^2}|\boldsymbol{x}-\hat{\boldsymbol{w}}|^2\right) \end{aligned} \tag{8}$$

となる．

3.2 ハイパーパラメータ最適化

経験ベイズ法では，ハイパーパラメータ a を周辺尤度 (5) に基づいて決定する．本論では周辺尤度を最大化する方法を用いる．周辺尤度の逆数の対数である確率的複雑さは，$\boldsymbol{w}^* = \boldsymbol{0}$ の場合の真の分布のエントロピー S を用いて

$$\begin{aligned} F(\boldsymbol{X}^n;a) &= -\log Z(\boldsymbol{X}^n;a) \\ &= n\cdot S + \frac{K}{2}\log(na^2+1) - \frac{n^2a^2|\boldsymbol{m}|^2}{2(na^2+1)} \end{aligned} \tag{9}$$

と表される．これを a^2 で微分すると

$$\frac{\partial F(\boldsymbol{X}^n;a)}{\partial\left(a^2\right)} = \frac{Kn^2\left(a^2 - \left(|\boldsymbol{m}|^2/K - n^{-1}\right)\right)}{2\left(na^2+1\right)^2}$$

となり，従って極値条件と $a^2 > 0$ から，ハイパーパラメータの最尤推定量として

$$\hat{a}^{-2}\left(\boldsymbol{X}^n\right) = \begin{cases} \infty & \text{if } |\boldsymbol{m}| \le \sqrt{K/n} \\ \frac{nK}{n|\boldsymbol{m}|^2 - K} & \text{if } |\boldsymbol{m}| > \sqrt{K/n} \end{cases} \tag{10}$$

が得られる．経験ベイズ法による予測分布の平均と分散は，(6) と (7) に (10) を代入することによって得られる．

4 汎化誤差

ベイズ汎化誤差は真の分布 (1) からみた予測分布 (8) の KL 擬距離

$$G(\boldsymbol{X}^n;K,\boldsymbol{w}^*) = \int q(\boldsymbol{x})\log\frac{q(\boldsymbol{x})}{p\left(\boldsymbol{x}|\boldsymbol{X}^n;\hat{a}\left(\boldsymbol{X}^n\right)\right)}d\boldsymbol{x} \tag{11}$$

として定義される．ここで，真のパラメータが $\boldsymbol{w}_0 = O(1)$ を用いて

$$\boldsymbol{w}^* = \boldsymbol{w}_0/\sqrt{n} \tag{12}$$

と書けると仮定する．このスケーリングを用いることにより，特異点 $\left\{(a,\boldsymbol{b}) \in R^1 \times R^K; a\boldsymbol{b} = \boldsymbol{0}\right\}$ に対応するモ

[1] 最も典型的な経験ベイズ法は，事前分布にハイパーパラメータを導入することによって適用される．式 (3) および (4) で表現されるモデルは，変数変換 $a\boldsymbol{b} \to \boldsymbol{w}$ により，

$$p(\boldsymbol{x}|\boldsymbol{w}) = \frac{1}{(2\pi)^{K/2}} \exp\left(-\frac{|\boldsymbol{x}-\boldsymbol{w}|^2}{2}\right)$$

$$\phi(\boldsymbol{w};a) = \frac{1}{(2\pi a^2)^{K/2}} \exp\left(-\frac{|\boldsymbol{w}|^2}{2a^2}\right)$$

となるので，K 次元正規分布モデルにおいて，平均ベクトル $\boldsymbol{w}$ の事前分布の標準偏差 a をハイパーパラメータとしたモデルと等価である．

デルから見た真の分布のカルバック擬距離が n^{-1} に比例し，モデル選択において最も重要な領域を連続的に解析できる [10]．経験ベイズ法の汎化誤差について，以下の定理が成り立つ．

定理 1 真のモデル (1) を，学習モデル (3)，(4) を用いて，ハイパーパラメータを周辺尤度に基づいて最尤推定する経験ベイズ学習を行ったときの汎化誤差の期待値は，

$$G(n;K,\boldsymbol{w}_0) = E_{\boldsymbol{X}^n}[G(\boldsymbol{X}^n;K,\boldsymbol{w}_0)] = \frac{\lambda(K,\boldsymbol{w}_0)}{n} + O\left(n^{-2}\right) \tag{13}$$

と漸近展開できる．ここで

$$2\lambda(K,\boldsymbol{w}_0) = |\boldsymbol{w}_0|^2 + E_{\boldsymbol{g}}\Big[I\Big(|\boldsymbol{w}_0+\boldsymbol{g}|>\sqrt{K}\Big)\, L\left(\boldsymbol{g};K,\boldsymbol{w}_0\right)\Big] \tag{14}$$

$$L\left(\boldsymbol{g};K,\boldsymbol{w}_0\right) = \frac{\left(|\boldsymbol{w}_0+\boldsymbol{g}|^2-K\right)^2 - 2\left(|\boldsymbol{w}_0+\boldsymbol{g}|^2-K\right)\boldsymbol{w}_0^t\left(\boldsymbol{w}_0+\boldsymbol{g}\right)}{|\boldsymbol{w}_0+\boldsymbol{g}|^2} \tag{15}$$

である．$\boldsymbol{g}$ は K 次元標準正規分布に従う確率変数であり，$E_{\boldsymbol{g}}[\cdot]$ は $\boldsymbol{g}$ に関する期待値である．また，$I(condition)$ は，$condition$ が真の時 1，偽の時 0 となる関数である．なお，$\lambda(K,\boldsymbol{w}_0)$ はベイズ学習における学習係数と呼ばれる．

（証明）

(11) に (1) と (8) を代入すると，

$$\begin{aligned} G(\boldsymbol{X}^n;K,\boldsymbol{w}^*) &= \int \frac{1}{(2\pi)^{K/2}} \exp\left(-\frac{|\boldsymbol{x}-\boldsymbol{w}^*|^2}{2}\right) \\ &\quad \cdot\left(-\frac{|\boldsymbol{x}-\boldsymbol{w}^*|^2}{2} + \frac{K}{2}\log\hat{\sigma}^2 + \frac{|\boldsymbol{x}-\hat{\boldsymbol{w}}|^2}{2\hat{\sigma}^2}\right) d\boldsymbol{x} \\ &= -\frac{K}{2} + \frac{K}{2}\log\hat{\sigma}^2 + \frac{K}{2\hat{\sigma}^2} + \frac{|\boldsymbol{w}^*-\hat{\boldsymbol{w}}|^2}{2\hat{\sigma}^2} \end{aligned}$$

となり，これに (6) および (7) を代入して

$$G(\boldsymbol{X}^n;K,\boldsymbol{w}^*) = -\frac{K}{2} + \frac{K}{2}\log\left(1+\frac{1}{n+\hat{a}^{-2}}\right) + \frac{(K+|\boldsymbol{w}^*|^2)\left(n+\hat{a}^{-2}\right)^2 - 2n\boldsymbol{w}^{*t}\boldsymbol{m}\left(n+\hat{a}^{-2}\right) + n^2|\boldsymbol{m}|^2}{2\left(n+\hat{a}^{-2}+1\right)\left(n+\hat{a}^{-2}\right)}$$

が得られる．ここで，式 (12) のスケーリングを適用する．この時，平均ベクトル $\boldsymbol{m}$ から

$$\boldsymbol{m} = \left(\boldsymbol{w}_0+\boldsymbol{g}\right)/\sqrt{n} \tag{16}$$

によって $\boldsymbol{g}$ を定めると，$\boldsymbol{g}$ は K 次元標準正規分布に従う確率変数になる．また，経験ベイズ法では $\hat{a}$ の値はサンプル $\boldsymbol{X}^n$ に依存して式 (10) で決定されるが，この値に拠らず $\left(n+\hat{a}^{-2}\right)^{-1}$ は $O(n^{-1})$ 以下であることを用いて，

$$G(\boldsymbol{X}^n;K,\boldsymbol{w}_0) = \frac{1}{2}\left\{\frac{|\boldsymbol{w}_0|^2}{n} - \frac{2\boldsymbol{w}_0^t\left(\boldsymbol{w}_0+\boldsymbol{g}\right)}{\left(n+\hat{a}^{-2}\right)} - \frac{n\,|\boldsymbol{w}_0+\boldsymbol{g}|^2}{\left(n+\hat{a}^{-2}\right)^2}\right\} + O\left(n^{-2}\right) \tag{17}$$

が成立する．以下，式 (10) のハイパーパラメータ決定ルールに基づいて，場合分けを行って検討する．

1. $|\boldsymbol{w}_0+\boldsymbol{g}| \le \sqrt{K}$ のとき

   $\hat{a}^{-2} = \infty$ であり，このとき式 (17) は

   $$G(\boldsymbol{X}^n;K,\boldsymbol{w}_0) = \frac{|\boldsymbol{w}_0|^2}{2n} + O(n^{-2}) \tag{18}$$

   となる．

2. $|\boldsymbol{w}_0+\boldsymbol{g}| > \sqrt{K}$ のとき

   $$\left(n+\hat{a}^{-2}\right)^{-1} = \left(1-\frac{K}{|\boldsymbol{w}_0+\boldsymbol{g}|^2}\right)n^{-1}$$

   であり，これを式 (17) に代入して整理すると，

   $$G(\boldsymbol{X}^n;K,\boldsymbol{w}_0) = \frac{1}{2n}\left\{|\boldsymbol{w}_0|^2 + L\left(\boldsymbol{g};K,\boldsymbol{w}_0\right)\right\} + O\left(n^{-2}\right) \tag{19}$$

   が得られる．

従って 1. の場合と 2. の場合とを合わせた汎化誤差の期待値は，式 (18) と式 (19) から

$$E_{\boldsymbol{X}^n}\left[G(\boldsymbol{X}^n;K,\boldsymbol{w}_0)\right] = O\left(n^{-2}\right) + \frac{1}{2n}\Big\{|\boldsymbol{w}_0|^2 + E_{\boldsymbol{g}}\Big[I\Big(|\boldsymbol{w}_0+\boldsymbol{g}|>\sqrt{K}\Big)L\left(\boldsymbol{g};K,\boldsymbol{w}_0\right)\Big]\Big\}$$

となり，定理 1 が証明された．

（証明終わり）

式 (14) は，$\boldsymbol{w}_0 = \boldsymbol{0}$ の場合には正規分布の期待値の積分を行って，

$$\begin{aligned} 2\lambda(K,\boldsymbol{0}) &= \int_{|\boldsymbol{g}|>\sqrt{K}} \left(\frac{1}{2\pi}\right)^{\frac{K}{2}} e^{-\frac{|\boldsymbol{g}|^2}{2}} \frac{\left(|\boldsymbol{g}|^2-K\right)^2}{|\boldsymbol{g}|^2} d\boldsymbol{g} \\ &= \frac{2\pi^{\frac{K}{2}}}{\Gamma\left(\frac{K}{2}\right)}\left(\frac{1}{2\pi}\right)^{\frac{K}{2}} \int_{\sqrt{K}}^{\infty} e^{-\frac{r^2}{2}}\left(r^2-K\right)^2 r^{K-3} dr \\ &= K\cdot\frac{\Gamma\left(\frac{K-2}{2},\frac{K}{2}\right)}{\Gamma\left(\frac{K}{2}\right)} \end{aligned} \tag{20}$$

と書ける．ここで第 2 種不完全ガンマ関数

$$\Gamma\left(z,P\right) = \int_P^{\infty} t^{z-1}e^{-t}dt$$

とガンマ関数 $\Gamma(z)=\Gamma(z,0)$ を用いた．

しかし一般の $\boldsymbol{w}_0$ の場合においては，(14) はこれ以上簡単な形には書けない．そこで，一般性を失うことなく $\boldsymbol{w}_0=(w_0,0,\cdots,0)^t$ とし，(14) を数値的に計算した．図 1 に $K=1,\cdots,6$ の場合の結果を示す．横軸に $w_0=|\boldsymbol{w}_0|=\sqrt{n}|\boldsymbol{w}^*|$，縦軸に $2\lambda/K$ をとった．正則モデルの場合の学習係数 $2\lambda=K$ も，比較のため図中に示した．図 1 から，$K\geq 5$ では $\boldsymbol{w}_0$ の値によらず，常に階層モデルの経験ベイズ法が正則モデルに優っていることがわかる．このことをもう少し精密に調べるために，以下で汎化誤差の $|\boldsymbol{w}_0|\to\infty$ における漸近展開を行う．

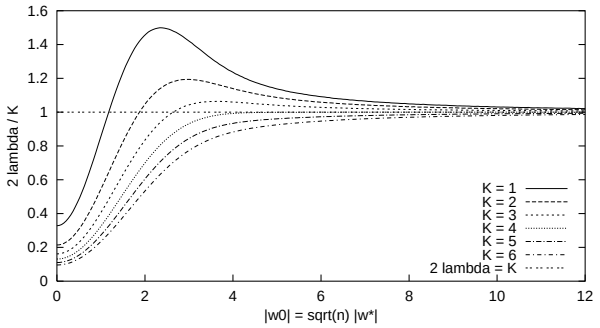

図 1: 汎化誤差

4.1 $|\boldsymbol{w}_0|\to\infty$ における漸近展開

式 (14) は

$$2\lambda(K,\boldsymbol{w}_0)=|\boldsymbol{w}_0|^2+E_{\boldsymbol{g}}[L(\boldsymbol{g};K,\boldsymbol{w}_0)] -E_{\boldsymbol{g}}\Big[I\Big(|\boldsymbol{w}_0+\boldsymbol{g}|\leq\sqrt{K}\Big)L(\boldsymbol{g};K,\boldsymbol{w}_0)\Big] \quad (21)$$

と書けるが，第 3 項は $|\boldsymbol{w}_0|\to\infty$ で $O\left(\exp\left(-|\boldsymbol{w}_0|^2\right)\right)$ で収束する．また，(15) を展開すると，

$$L(\boldsymbol{g};K,\boldsymbol{w}_0)=-|\boldsymbol{w}_0|^2+|\boldsymbol{g}|^2 +\frac{1}{|\boldsymbol{w}_0|^2}\left\{K^2-2K\left(2\boldsymbol{w}_0^t\boldsymbol{g}+|\boldsymbol{g}|^2\right)+2K\frac{\left(2\boldsymbol{w}_0^t\boldsymbol{g}+|\boldsymbol{g}|^2\right)^2}{|\boldsymbol{w}_0|^2} +2K\boldsymbol{w}_0^t\boldsymbol{g}-2K\boldsymbol{w}_0^t\boldsymbol{g}\frac{2\boldsymbol{w}_0^t\boldsymbol{g}+|\boldsymbol{g}|^2}{|\boldsymbol{w}_0|^2}\right\}+o\left(|\boldsymbol{w}_0|^{-2}\right) \quad (22)$$

となるので，これを (21) に代入して，$E_{\boldsymbol{g}}\left[|\boldsymbol{g}|^2\right]=K$，$E_{\boldsymbol{g}}[\boldsymbol{w}_0^t\boldsymbol{g}]=0$ および $E_{\boldsymbol{g}}\left[(\boldsymbol{w}_0^t\boldsymbol{g})^2\right]=|\boldsymbol{w}_0|^2$ を用いると，

$$2\lambda(K,\boldsymbol{w}_0)=K-\frac{K(K-1)}{|\boldsymbol{w}_0|^2}+o\left(|\boldsymbol{w}_0|^{-2}\right) \quad (23)$$

が得られる．$K=4$ をさかいに $|\boldsymbol{w}_0|^{-2}$ のオーダーの符号が反転し，漸近的な振る舞いが変化することがわかる．

5 学習誤差

ベイズ学習誤差は

$$T(\boldsymbol{X}^n;K,\boldsymbol{w}^*) \;=\frac{1}{n}\sum_{i=1}^{n}\log\frac{q(\boldsymbol{x}_i)}{p(\boldsymbol{x}_i|\boldsymbol{X}^n;\hat{a}\,(\boldsymbol{X}^n))} \quad (24)$$

で定義される．以下の定理が成り立つ．

定理 2 真のモデル (1) を，学習モデル (3)，(4) を用いてハイパーパラメータを周辺尤度に基づいて最尤推定する経験ベイズ学習を行ったときの学習誤差の期待値は，

$$T(n;K,\boldsymbol{w}_0)=E_{\boldsymbol{X}^n}[T(\boldsymbol{X}^n;K,\boldsymbol{w}_0)] =\frac{\mu(K,\boldsymbol{w}_0)}{n}+O\left(n^{-2}\right) \quad (25)$$

と漸近展開できる．ここで

$$2\mu(K,\boldsymbol{w}_0)=|\boldsymbol{w}_0|^2 +E_{\boldsymbol{g}}\Big[I\Big(|\boldsymbol{w}_0+\boldsymbol{g}|>\sqrt{K}\Big)M(\boldsymbol{g};K,\boldsymbol{w}_0)\Big] \quad (26)$$

$$M(\boldsymbol{g};K,\boldsymbol{w}_0)=-|\boldsymbol{w}_0+\boldsymbol{g}|^2+\frac{K^2}{|\boldsymbol{w}_0+\boldsymbol{g}|^2} \quad (27)$$

である．

（証明）

(24) に (1) と (8) を代入すると，

$$\begin{aligned}T(\boldsymbol{X}^n;K,\boldsymbol{w}^*)&=\frac{1}{n}\sum_{i=1}^{n}\left(-\frac{|\boldsymbol{x}_i-\boldsymbol{w}^*|^2}{2}+K\log\hat{\sigma}+\frac{|\boldsymbol{x}_i-\hat{\boldsymbol{w}}|^2}{2\hat{\sigma}^2}\right)\\ &=K\log\hat{\sigma}-\frac{1}{2n}\left(1-\frac{1}{\hat{\sigma}^2}\right)\sum_{i=1}^{n}|\boldsymbol{x}_i-\boldsymbol{m}|^2\\ &\quad-\frac{1}{2}\left(|\boldsymbol{m}-\boldsymbol{w}^*|^2-\frac{1}{\hat{\sigma}^2}|\boldsymbol{m}-\hat{\boldsymbol{w}}|^2\right)\\ &=K\log\hat{\sigma}-\frac{S_0}{2}\left(1-\frac{1}{\hat{\sigma}^2}\right)-\frac{S_1}{2n}+\frac{1}{2\hat{\sigma}^2}|\boldsymbol{m}-\hat{\boldsymbol{w}}|^2\end{aligned}$$

となる．ただし，$S_0(\boldsymbol{X}^n)=\sum|\boldsymbol{x}_i-\boldsymbol{m}|^2/n$ および $S_1(\boldsymbol{X}^n)=n|\boldsymbol{m}-\boldsymbol{w}^*|^2$ を用いた．これに (6) および (7) を代入して

$$T(\boldsymbol{X}^n;K,\boldsymbol{w}^*)=\frac{K}{2}\log\left(1+\frac{1}{n+\hat{a}^{-2}}\right)-\frac{S_0}{2(n+\hat{a}^{-2}+1)} -\frac{S_1}{2n}+\frac{\hat{a}^{-4}|\boldsymbol{m}|^2}{2\,(n+\hat{a}^{-2}+1)\,(n+\hat{a}^{-2})} \quad (28)$$

を得る．式 (12) のスケーリングを適用し，(16) も用いて展開すると，

$$T(\boldsymbol{X}^n;K,\boldsymbol{w}_0)=-\frac{S_1}{2n}+\frac{(K-S_0)}{2(n+\hat{a}^{-2})}+\frac{\hat{a}^{-4}|\boldsymbol{w}_0+\boldsymbol{g}|^2}{2n(n+\hat{a}^{-2})^2} +O\left(n^{-2}\right) \quad (29)$$

となる．(29) で，サンプル $\boldsymbol{X}^n$ に依存する変数は S_0, S_1 および $\hat{a}$ であるが，$\hat{a}$ の値は式 (10) より，平均値の最尤推定量 $\boldsymbol{m}$ のみに依存する．一方，S_0 は分散の最尤推

定量であるため，$\hat{a}$ と S_0 は独立である．従って (29) の期待値は

$$
\begin{aligned}
T(n;K,\boldsymbol{w}_0) &= E_{\boldsymbol{X}^n}[T(\boldsymbol{X}^n;K,\boldsymbol{w}_0)] \\
&= E_{\boldsymbol{X}^n}\left[-\frac{S_1}{2n}\right] + E_{\boldsymbol{X}^n}[K-S_0]\cdot E_{\boldsymbol{X}^n}\left[\frac{1}{2(n+\hat{a}^{-2})}\right] \\
&\qquad + E_{\boldsymbol{X}^n}\left[\frac{\hat{a}^{-4}|\boldsymbol{w}_0+\boldsymbol{g}|^2}{2n(n+\hat{a}^{-2})^2}\right] + O\big(n^{-2}\big) \\
&= \frac{|\boldsymbol{w}_0|^2}{2n} + E_{\boldsymbol{X}^n}[H\left(\boldsymbol{X}^n\right)] + O\big(n^{-2}\big) \qquad (30)
\end{aligned}
$$

となる．ここで $E_{\boldsymbol{X}^n}[S_0] = (1-1/n)\,K$ 及び $E_{\boldsymbol{X}^n}[S_1] = K$ を用い，また

$$
H\left(\boldsymbol{X}^n\right) = -\frac{|\boldsymbol{w}_0+\boldsymbol{g}|^2}{2n} + \frac{\hat{a}^{-4}|\boldsymbol{w}_0+\boldsymbol{g}|^2}{2n\left(n+\hat{a}^{-2}\right)^2} \qquad (31)
$$

とおいた．(31) の値について，4 節と同様に式 (10) に基づいた場合分けを行う．

1. $|\boldsymbol{w}_0+\boldsymbol{g}| \le \sqrt{K}$ のとき

   $\hat{a}^{-2} = \infty$ であり，このとき (31) は

   $$
   H\left(\boldsymbol{X}^n\right) = 0 \qquad (32)
   $$

   となる．

2. $|\boldsymbol{w}_0+\boldsymbol{g}| > \sqrt{K}$ のとき

   (31) に $\hat{a}^{-2} = nK/\left(|\boldsymbol{w}_0+\boldsymbol{g}|^2 - K\right)$ を代入すると，

   $$
   H\left(\boldsymbol{X}^n\right) = \frac{1}{2n}\left\{-|\boldsymbol{w}_0+\boldsymbol{g}|^2 + \frac{K^2}{|\boldsymbol{w}_0+\boldsymbol{g}|^2}\right\} + O\big(n^{-2}\big) \qquad (33)
   $$

   である．

従って 1. の場合と 2. の場合とを合わせた学習誤差の期待値は，(32)，(33) 及び (30) より

$$
\begin{aligned}
&T(n;K,\boldsymbol{w}_0) = O\left(n^{-2}\right) \\
&+ \frac{1}{2n}\left\{|\boldsymbol{w}_0|^2 + E_{\boldsymbol{g}}\left[I\Big(|\boldsymbol{w}_0+\boldsymbol{g}| > \sqrt{K}\Big) M\left(\boldsymbol{g};K,\boldsymbol{w}_0\right)\right]\right\}
\end{aligned}
$$

となり，定理 2 が証明された．

（証明終わり）

汎化誤差の場合と同様に，(26) は $\boldsymbol{w}_0 = \boldsymbol{0}$ の場合のみより簡単に書ける．

$$
2\mu(K,\boldsymbol{0}) = 2\lambda(K,\boldsymbol{0}) - K\cdot\frac{2}{\Gamma\left(\frac{K}{2}\right)}\left(\frac{K}{2}\right)^{\frac{K-2}{2}} e^{-\frac{K}{2}}
$$

ただし，(20) の $\lambda(K,\boldsymbol{0})$ を用いた．一般の $\boldsymbol{w}_0$ の場合について (26) を数値的に計算した結果を図 2 に示す．正則モデルの場合の学習誤差の主要項の係数は $2\mu = -K$ である．図 1 と見比べて分かるように，汎化誤差と学習誤差は正則モデルの場合のように対称的にはならない．

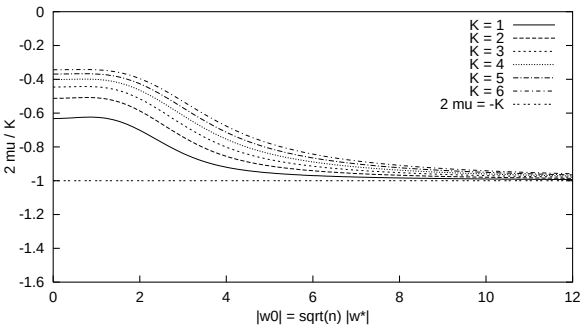

図 2: 学習誤差

5.1 学習誤差の $|\boldsymbol{w}_0| \to \infty$ における漸近展開

式 (26) を，汎化誤差の場合と同様に漸近展開すると，

$$
2\mu(K,\boldsymbol{w}_0) = -K + \frac{K^2}{|\boldsymbol{w}_0|^2} + o\left(|\boldsymbol{w}_0|^{-2}\right) \qquad (34)
$$

が得られる．$|\boldsymbol{w}_0|^{-2}$ のオーダーの符号は K に拠らず常に正である．

6 考察

真の分布 (1) を学習モデル (2) でベイズ学習した場合（a，$\boldsymbol{b}$ ともにパラメータ）の汎化誤差はすでに解析されている [10] [2]．この結果に対し，本論で解析した経験ベイズ法の汎化誤差を比較したものを図 3 に示す．太線が経験ベイズ法，細線がベイズ法の場合であり，それぞれ $K = 2,3,4$ の場合を示した．各次元において $1 < |\boldsymbol{w}_0| < 2$ の領域で，ベイズ法と経験ベイズ法との優劣関係が入れ替わっていることが分かる．

本論の解析結果から以下のことが分かる．まず第 1 に，経験ベイズ法でも特異点の過学習抑制効果は得られ，高次元で Stein 推定量に似た性質が得られる．ただし，真の分布のパラメータ $\boldsymbol{w}^*$ に拠らず常に正則モデルを優越する次元が，経験ベイズ法では $K \ge 5$ であるのに対し，ベイズ推定では $K \ge 4$ であるという違いがある[3]．また，$\boldsymbol{w}^* = \boldsymbol{0}$ の場合の汎化誤差にも違いが見られ，経験ベイズ法では (20) で表されるように K に依存するのに対し，ベイズ推定では $2\lambda = 1$ となり，K によらず一定である．

両者の類似性を利用した工学的応用も考えられる．ベイズ推定は，すべてのパラメータを周辺化して予測分布を求める方法であるのに対し，本論で解析した経験ベイズ法は，ある階層のパラメータについてのみ周辺化し，残りの階層のパラメータをハイパーパラメータとみな

[2] [10] では回帰問題を取り扱っているが，主要項についてはこのモデルの場合と変わらないことが確認されている．

[3] ハイパーパラメータを求める方法として，周辺尤度最大化以外の方法を用いることができ，ある方法を用いると，正則モデルを優越する次元を $K \ge 3$ にできる．この時，経験ベイズ推定量は James-Stein 推定量となっている [15].

して周辺尤度に基づいて最尤推定を行う方法である．本論の結果から，この方法でも特異点が過学習を抑制する効果が得られることが分かる．階層モデルでベイズ学習を行う場合，事後分布に基づいて MCMC 法などでパラメータ空間上にサンプルを発生させて予測分布を得る方法が知られており，広く用いられている．階層モデルは各階層ごとに見れば（他の階層のパラメータを定数とみれば）正則モデルであるので，一部の階層のパラメータに関する積分は解析的に実行できる場合がある．経験ベイズ法では，残りの層について周辺尤度を最大化するだけで良いので，計算負荷を大幅に減らすことができる．また，解析的な積分ができない場合においても，サンプルを発生させる次元を減らすことができ，やはり計算量の点で有利である．

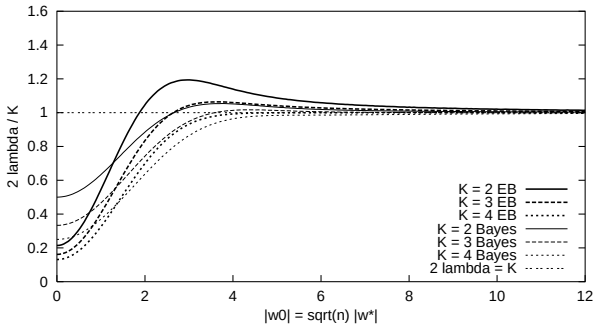

図 3: 汎化誤差（経験ベイズ法 (EB) vs ベイズ法）

7　まとめと今後の課題

本論では，最も簡単な階層型学習モデルを用いて経験ベイズ学習を行った場合の汎化誤差と学習誤差について，真のモデルと特異点に対応するモデルとのカルバック擬距離がサンプル数に反比例するようなスケーリングを用いて解析した．今後の課題は，回帰モデルの場合の解析，より実用的な階層モデルにおいて一部の階層のパラメータをハイパーパラメータとみなして経験ベイズ法を行った場合の汎化誤差を解明することなどである．

参考文献

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

[2] S. Watanabe, “Algebraic geometrical methods for hierarchical learning mahicnes,” *Neural Networks*, Vol.14, No.8, pp.1049–1060, 2001.

[3] 渡辺澄夫, データ学習アルゴリズム, 共立出版, 東京, 2001.

[4] K. Watanabe, S. Watanabe, “縮小ランク回帰モデルのベイズ汎化誤差について,” **電子情報通信学会誌**, Vol.J86-A, No.3, pp.278–287, 2003.

[5] 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.225, pp.13–18, 2004.

[6] K. Yamazaki, S. Watanabe, “Singularities in mixture models and upper bounds of stochastic complexity,” *Neural Networks*, Vol.16, No.7, pp.1029–1038, 2003.

[7] K. Yamazaki, S. Watanabe, “Stochastic complexity of Bayesian networks,” *Proc. of UAI*, pp.592–599, 2003.

[8] K. Yamazaki, S. Watanabe, “Stochatic complexities of hidden Markov models,” *Proc. of NNSP*, pp.179–188, 2003.

[9] 山崎啓介, 青柳美輝, 渡辺澄夫, “ニュートン図形を用いた確率的複雑さの解析法,” **信学技報**, Vol.104, No.225, pp.19–24, 2004.

[10] S. Watanabe, S. Amari, “Learning Coefficients of Layered Models When the True Distribution Mismatches the Singularities,” *Neural Computation*, Vol.15, pp.1013–1033, 2003.

[11] B. Efron, C. Morris, “Stein’s estimation rule and its competitors — an empirical Bayes approach,” *Journal of the Amrican Statistical Association*, Vol.68, pp.117–130, 1973.

[12] R. E. Kass, D. Steffey, “Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models),” *Journal of the Amrican Statistical Association*, Vol.84, pp.717–726, 1989.

[13] C. Stein, “Inadmissibility of the usual estimator for the mean of a multivariate normal distribution,” *Proc. of the 3rd Berkeley symp. on Mathematical Statistics and Probability*, Vol.1, pp.197–206, 1956, Berkeley: University of California.

[14] W. James, C. Stein, “Estimation with quadratic loss,” *Proc. of the 4th Berkeley symp. on Mathematical Statistics and Probability*, Vol.1, pp.361–379, 1961, Berkeley: University of California.

[15] 中島伸一, 渡辺澄夫, “ハイパーパラメータ最適化法における汎化誤差について,” **信学技報**, Vol.104, No.225, pp.7–12, 2004.

[16] 福水健次，栗木哲，竹内啓，赤平昌文, 特異モデルの統計学, 岩波書店, 東京, 2004.