

カルバック情報量の分割による特異モデルの確率的複雑さの計算法

永田 賢二[†] 渡辺 澄夫^{††}

[†] 東京工業大学 工学部 情報工学科 〒 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: [†]kenji.nagata@cs.pi.titech.ac.jp, ^{††}swatanab@pi.titech.ac.jp

あらまし 神経回路網、混合正規分布、ベイズネットワーク、隠れマルコフモデル、ボルツマンマシンなど、近年情報科学において広く用いられるようになった学習モデルの多くは特異モデルであることが知られている。これらのフィッシャー情報行列は正定値ではなく、従来の統計的漸近理論が成り立たず、確率的複雑さを計算することは容易ではない。本論ではカルバック情報量を分割し、MCMC法と代数幾何学の理論との組み合わせによって確率的複雑さを計算するアルゴリズムをいくつか提案し、それぞれのアルゴリズムにおいて、計算時間や計算精度を比較することにより、それらの性質を数値実験により明らかにする。

キーワード 特異モデル、カルバック情報量、確率的複雑さ

A Method to Estimate the Stochastic Complexity of Singular Learning Machines by Decomposition of Kullback Information

Kenji NAGATA[†] and Sumio WATANABE^{††}

[†] Department of Computer Science Tokyo Institute of Technology, 4259

Nagatsuda, Midori-ku, Yokohama, 226-8503 Japan

^{††} PI Lab., Tokyo Institute of Technology.

E-mail: [†]kenji.nagata@cs.pi.titech.ac.jp, ^{††}swatanab@pi.titech.ac.jp

Abstract A lot of learning machines such as neural networks, normal mixtures, Bayesian networks, and hidden Markov models are singular statistical models. Their Fisher information matrices are not positive definite, hence it is not easy to calculate the stochastic complexity. In this paper, we propose some new methods to calculate the stochastic complexity by decomposing the Kullback information. By comparing the calculating times and accuracies of each proposed method, the properties of them are shown by experimental results.

Key words Singular Learning Machines, Kullback Information, Stochastic Complexity

1. ま え が き

神経回路網、混合正規分布、ベイズネットワーク、隠れマルコフモデル、ボルツマンマシンなど、例からの学習によって確率的な推論方式を獲得するモデルは、階層的な構造や対称性に起因するフィッシャー情報行列の縮退が生じるため、統計的正則モデルではないことが知られている。これらの学習モデルは総称して特異モデルと呼ばれているが、特異モデルにおいては、最尤推定量の分布もベイズ事後分布も正規分布に漸近しないために、従来の統計学的あるいは統計物理学的な理論は成り立たず、最適な設計法が未だに確立されていない [1] [7] [8] [9] [10] [11] [12]。

学習理論においては、「例からの学習」についての数学的な理論と情報学的なアルゴリズムが研究されているが、その主要

な目的は次のふたつである。

(1) 順問題：サンプルを発生している真の確率分布 $q(x)$ 、真の分布を推測するために用いられる学習モデル $p(x|w)$ およびその事前分布 $\varphi(w)$ が与えられたとき、「学習結果と真の分布との違い」を解明すること。

(2) 逆問題：真の分布 $q(x)$ が不明であって、サンプルだけが与えられたときに、「学習結果と真の分布との違い」をできるだけ小さくするように、学習モデルや事前分布を定めるアルゴリズムを作ること。

このうち、情報科学への応用においては、(2)の逆問題が重要であるが、逆問題を考えるためには、そのための基礎となる(1)

の順問題を解決しておく必要がある。

統計的正則モデルにおいては、最尤推定量の漸近分布やベイズ事後分布の漸近形を容易に導くことができ、順問題について解明されているので、その結果に基づいて、逆問題へのアプローチとして AIC, BIC, MDL などの学習モデル選択のアルゴリズムが考案されている。

しかしながら、特異モデルにおいては、未だに順問題が解明されていないために、数学的に意味のある学習アルゴリズムはまだ確立されていない。そこで、本論では、特異モデルの順問題について、カルバック情報量を分割することにより、確率的複雑さを分解する方法を提案し、その有効性をいくつかの実験により明らかにする。

2. 特異モデルの学習理論

2.1 ベイズ学習の枠組

本論では、特異モデルの学習において、現在のところ、最も高精度な推測を実現すると考えられているベイズ法について検討する。

N 次元ユークリッド空間 R^N 上の確率変数 X が確率密度関数 $q(x)$ を持つとし、 X の独立なサンプルが n 個得られた場合を考える。そのサンプルを

$$X^n = (X_1, X_2, \dots, X_n)$$

と書く。サンプルは、それが観測される度に確率的に変動するのであるから、確率変数である。このサンプルから真の分布を推測するために d 次元ユークリッド空間 R^d 内に値を取るパラメータ w を持つ確率推論モデル $p(x|w)$ を考える。 $p(x|w)$ は、神経回路網、混合正規分布、ベイズネットワークなどの高度に複雑な学習モデルを表すものとする。またパラメータ w が従う確率分布(事前分布)を $\varphi(w)$ とする。このとき、サンプル X^n とパラメータ w の同時分布は

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

であるから、「サンプル X^n が与えられたという条件下におけるパラメータ w の分布」は、ベイズの定理によって

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

となる。ここで、 $Z(X^n)$ は正規化定数

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

であるが、これは、学習モデルと事前分布 $(p(x|w), \varphi(w))$ の尤度に等しい(学習モデルと事前分布の「証拠」と呼ばれることもあり、統計物理学における分配関数と同じである)。ベイズ学習においては、この事後分布で学習モデルを平均することにより、予測分布 $p(x|X^n)$ を構成する。すなわち、

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

この予測分布 $p(x|X^n)$ は、「サンプル X^n が与えられたというもとで X の密度関数を推測したもの」である。予測分布は真の分布に近いと思われるが、サンプルが有限であることおよびサンプルがゆらぎを持つことから、完全には一致しない。学習理論における最初の課題は、「予測分布と真の分布 $q(x)$ の違いは、サンプルが増えるにつれて、どのような早さで小さくなってゆくか」を明らかにすることである。

2.2 カルバック情報量

ここで、一般に二つの確率分布 $q(x)$ と $p(x)$ の違いを表す量としてカルバック情報量を導入する。それは

$$D(q||p) = \int q(x) \log \frac{q(x)}{p(x)} dx$$

によって定義される。一般に $D(q||p) \neq D(p||q)$ なので、カルバック情報量は、通常の意味での距離にはならないが、任意の $q(x), p(x)$ に対して

$$D(q||p) \geq 0$$

が成り立ち、また

$$D(q||p) = 0 \iff q(x) = p(x) \quad (\forall x)$$

が成り立つ。

さて、学習理論においては、次のふたつのカルバック情報量が重要である。ひとつは、真の分布と学習モデルとのカルバック情報量である。

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

もうひとつは、真の分布とベイズ予測分布とのカルバック情報量である。

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

ある。 $H(w)$ は真の分布と学習モデルが与えられれば、パラメータの関数として確定する。一方、 $G(X^n)$ は、サンプル X^n の出方によって確率的に変動する確率変数である。 $G(X^n)$ は「ベイズ予測分布が、どの程度真の分布に近いか」を表すので、汎化誤差と呼ばれている。カルバック情報量 $H(w)$ に基づいて、汎化誤差がどのような性質をもっているかを解明することが、学習理論における順問題である。

2.3 特異モデルの性質

特異モデルの学習理論によって、汎化誤差 $G(X^n)$ は、ある定数 λ が存在して、次のように表されることが知られている [11]。

$$E[G(X^n)] = \frac{\lambda}{n} + o\left(\frac{1}{n}\right).$$

ここで λ は学習係数と呼ばれる。学習における順問題は、学習係数 λ を求めることに帰着するが、これまでの研究によって、学習係数は、次の幾つの特徴づけができることが知られている。

(1) $(-\lambda)$ は、ゼータ関数

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

の最も原点に近い極に等しい [11].

(2) 確率的複雑さ $F(n)$ を

$$F(n) = -\log \int \exp(-nH(w)) \varphi(w) dw$$

と定義すると、 $F(n)$ は λ と定数 m を用いて次のように表される [10].

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

(3) λ は次の極限值である [12].

$$\lambda = \lim_{t \rightarrow 0} \frac{\log(V(at)/V(t))}{\log t}$$

ここで

$$V(t) = \int_{H(w) < t} \varphi(w) dw.$$

このうち、(1) は、カルバック情報量の特異点を解消することにより学習係数が求められることを述べており、様々な学習モデルの学習係数を理論的に求める場合に用いられる。また (2)(3) はカルバック情報量と事前分布が与えられた場合に、学習係数を数値計算で求める場合に利用される。

3. 提案方法

本論では、カルバック情報量が

$$H(w) = H_1(w) + H_2(w)$$

と分解されて、かつ $H_1(w)$ については確率的複雑さ $F_1(n)$ が解明されている場合に、 $H(w)$ の確率的複雑さ $F(n)$ を求めるアルゴリズムをいくつか提案し、それぞれのアルゴリズムについての計算時間や計算精度を比較することにより、それらの性質を明らかにする。まず、提案するアルゴリズムに先だって、その基礎となる定理を述べて証明を行う。

3.1 基礎定理と証明

確率分布 $\rho_n(w)$ を次のように定義する。

$$\rho_n(w) \propto \exp(-nH_1(w)) \varphi(w).$$

このとき、次の定理がなりたつ。

定理 1

$$G(n) = -\log \int \exp(-nH_2(w)) \rho_n(w) dw$$

と $G(n)$ を定義すると確率的複雑さについて次の関係が成り立つ。

$$F(n) = F_1(n) + G(n)$$

(定理 1 の証明)

$$Z(n) = \int e^{-nH(w)} \varphi(w) dw$$

とおくと、確率的複雑さの定義から、

$$F(n) = -\log Z(n)$$

であるが、これに関係式

$$Z(n) = \int e^{-nH_1(w)} \varphi(w) dw \times \frac{\int e^{-nH_2(w)-nH_1(w)} \varphi(w) dw}{\int e^{-nH_1(w)} \varphi(w) dw}$$

を代入すると定理が得られる (定理 1 の証明終わり)。

3.2 計算アルゴリズム

上記の定理より、 $G(n)$ を数値的に求めることで、確率的複雑さ $F(n)$ を計算できることがわかる。そこで、 $G(n)$ を計算するアルゴリズムをいくつか提案する。まず、提案する手法に先だって、MCMC 法の一つであるメトロポリス法について説明する。

メトロポリス法

ある確率分布 $p(w)$ に法則収束する $\{w_k\}$ は次のようにして作り出せる。

1. w の初期値を設定する。
2. w に乱数を加えることにより w' を得る。
3. もしも $H_1(w) > H_1(w')$ であれば、 w' を採択し、2. の手続きにもどる。
4. もしも $H_1(w) \leq H_1(w')$ であれば、確率

$$P = \exp(-nH_1(w') + nH_1(w))$$

で w' を採択し、確率 $1-P$ で w を採択し、2. の手続きに戻る。

この手法に基づいて、 $F(n)$ を計算するアルゴリズムを3つ提案する。

計算アルゴリズム 1

$G(n)$ に含まれる積分を確率分布 $\rho_n(w)$ による $\exp(-nH_2(w))$ の平均値とみなし、 $G(n)$ を計算する。具体的なアルゴリズムを以下に示す。

1. $\rho_n(w)$ に従うサンプル $\{w_k; k = 1, 2, \dots, K\}$ を取り出す。
- 2、

$$y(n) = -\log \left\{ \frac{1}{K} \sum_{k=1}^K \exp(-nH_2(w_k)) \right\}$$

を計算する。

3. $F_1(n) + y(n)$ を目的の値とする。

計算アルゴリズム 2

$$g(n, t) = -\log \int \exp(-ntH_2(w))\rho_n(w)dw$$

と $g(n, t)$ を定義する。すると $G(n)$ は以下のようにして表せる。

$$\begin{aligned} G(n) &= g(n, 1) \\ &= \int_0^1 \frac{dg}{dt} dt \\ &= \int_0^1 dt \frac{\int nH_2(w) \exp(-ntH_2(w))\rho_n(w)dw}{\int \exp(-ntH_2(w))\rho_n(w)dw} \end{aligned}$$

これは、確率分布 $\exp(-ntH_2(w))\rho_n(w)$ による $nH_2(w)$ の平均値とみなすことができる。そこで、以下のようなアルゴリズムを提案する。

- 1、 t についての積分区間 $[0, 1]$ を細かく分割し、各 t について
- 2、3の手続きを行う。
- 2、 $\exp(-ntH_2(w))\rho_n(w)$ に従うサンプル $\{w_k; k = 1, 2, \dots, K\}$ を取り出す。
- 3、

$$y(n, t) = \frac{1}{K} \sum_{k=1}^K nH_2(w_k)$$

を計算する。

- 4、 $F_1(n) + \sum_t y(n, t)$ を目的の値とする。

計算アルゴリズム3

ここで計算するものは、アルゴリズム2と同じく $g(n, 1)$ を計算する。そこで以下のように計算する。

- 1、 $\rho_n(w)$ に従うサンプル $\{w_k; k = 1, 2, \dots, K\}$ を取り出す。
- 2、 t についての積分区間 $[0, 1]$ を細かく分割し、各 t について $y_1(n, t)$ 、 $y_2(n, t)$ を以下のように計算する。

$$y_1(n, t) = \frac{1}{K} \sum_{k=1}^K nH_2(w_k) \exp(-ntH_2(w_k))$$

$$y_2(n, t) = \frac{1}{K} \sum_{k=1}^K \exp(-ntH_2(w_k))$$

- 3、 $F_1(n) + \sum_t \frac{y_1(n, t)}{y_2(n, t)}$ を目的の値とする。

4. 実験

提案するアルゴリズムの評価を行うため、また分割することの有効性を評価するために、以下のようなモデルで実験を行った。本論で提案した方法は、確率的複雑さが未知であるカルバック情報量に適用することが目的であるが、以下では提案する方法の有効性を確認するために確率的複雑さが理論的に解明できるモデルについて調べることにする。

今回用いたモデルは、パラメータ次元 $d = 2$ とし、以下のようなカルバック情報量、事前分布について実験を行った。

$$H(a, b) = H_1(a, b) + H_2(a, b)$$

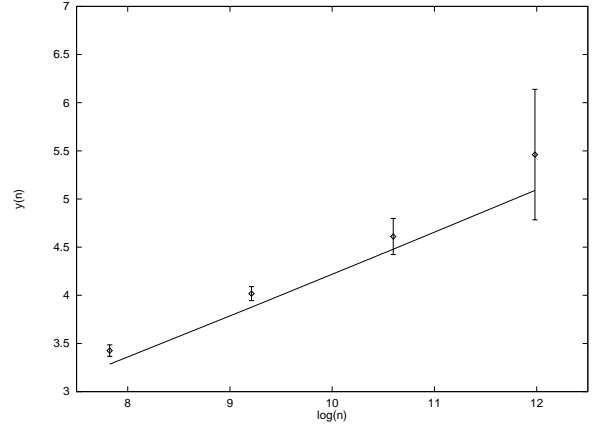


図1 アルゴリズム1の結果

$$H_1(a, b) = b^4$$

$$H_2(a, b) = a^2b^2$$

$$\varphi(a, b) = \frac{1}{2\pi} \exp\left(-\frac{a^2 + b^2}{2}\right)$$

このとき、 $H(a, b)$ 、 $H_1(a, b)$ に対する確率的複雑さは理論的に積分計算を行うことができ、以下ようになる。

$$F(n) = -\log \int da \int db \exp(-nH(a, b))\varphi(a, b)$$

$$\begin{aligned} &= \frac{1}{2} \log n - \log \log n \\ &\quad + \frac{1}{2} \log \pi + \log 4 - \frac{4C}{\log n} + O\left(\frac{1}{\sqrt{n}}\right) \end{aligned}$$

$$\begin{aligned} C &= \int_1^\infty \frac{1}{b} \exp(-b^4) db \\ &\quad + \int_0^1 \frac{1}{b} (\exp(-b^4) - 1) db + \frac{3}{2} \log 2 \end{aligned}$$

$$F_1(n) = -\log \int da \int db \exp(-nH_1(a, b))\varphi(a, b)$$

$$= \frac{1}{4} \log n + \frac{1}{2} \log \pi - \frac{1}{2} \log 2 - \log C_1$$

$$C_1 = \int_0^\infty \exp(-b^4) db$$

以下では、この値を理論値として実験を行なう。

4.1 実験 1

提案する3つのアルゴリズムの評価を行うため、それぞれのアルゴリズムに対し上記のモデルで実験を行った。ここで、実験の条件として、メトロポリス法のサンプル生成個数（以下繰り返し回数とする）を $K = 300000$ とし、各パラメータの更新時の乱数の上限（以下ばらつきとする）を $O\left(\frac{1}{\sqrt{n}}\right)$ とした。また、アルゴリズム2、3について、 t についての積分区間の分割は、 $[0, 0.1]$ の区間を $\frac{1}{10000}$ ずつ、 $[0.1, 1]$ の区間を $\frac{1}{1000}$ ずつ分割するように設定した。

各アルゴリズムについて n をいくつか設定し、各 n について

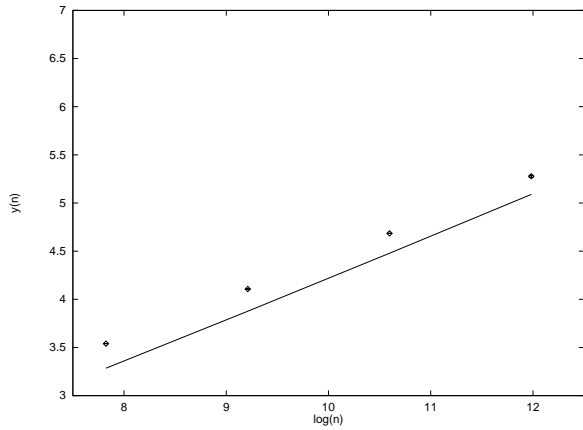


図 2 アルゴリズム 2 の結果

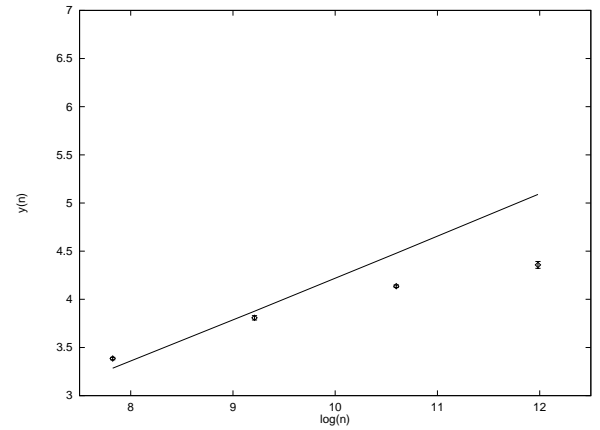


図 4 従来法の結果

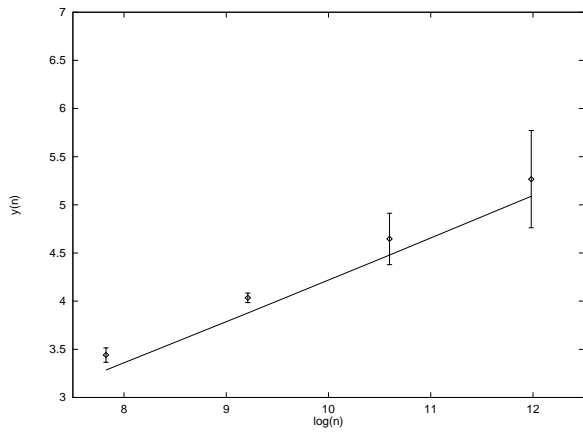


図 3 アルゴリズム 3 の結果

実験を 10 回行った結果を図 1、図 2、図 3 に示す。このグラフは横軸に $\log n$ 、縦軸に $F(n)$ をとったものであり、理論値を実線で表し、10 回行った結果の平均と分散を「(平均)±(標準偏差の二倍)」で表したものである。また、各アルゴリズムにおいて、 $F(n)$ を一回計算するためにかかった時間の平均を表 1 に示す。

以上の結果を見ると、どのアルゴリズムも理論値と比較するとわずかに大きく誤差が生じていることがわかる。分散は、アルゴリズム 1、3 では n が大きくなるにつれて大きくなっているのに対し、アルゴリズム 2 ではほぼ一定になることがわかった。また計算時間はアルゴリズム 1 では約 4 秒、アルゴリズム 2 では約 3 時間、アルゴリズム 3 では約 15 分かかった。計算時間に差が生じた理由は、アルゴリズム 1 と 2 では目的の値をひとつ計算するのにサンプル系列が 1 では 1 系列、2 では約 2 千系列必要とするからであり、1 と 3 では必要とするサンプル系列は同じであるが、3 では目的の値を計算するのにパラメータに関する積分計算を 2 つ行うためである。

4.2 実験 2

次に、分割することの有効性を評価するため、分割せずに $F(n)$ を計算する従来法との比較を行った。従来の方はアル

表 1 $F(n)$ を一回計算するのにかかる時間 (秒)

	計算時間
アルゴリズム 1	4.025
アルゴリズム 2	8535.850
アルゴリズム 3	860.700
従来法	8998.325

ゴリズム 2 において $G(n)$ を求めるかわりに $F(n)$ を計算するものである。具体的には、

$$f(n, t) = -\log \int \exp(-ntH(w))\varphi(w)dw$$

と $f(n, t)$ を定義し、

$$\begin{aligned} F(n) &= f(n, 1) \\ &= \int_0^1 \frac{df}{dt} dt \\ &= \int_0^1 \frac{\int nH(w) \exp(-ntH(w)) \varphi(w) dw}{\int \exp(-ntH(w)) \varphi(w) dw} \end{aligned}$$

として $F(n)$ を計算する方法である。具体的なアルゴリズムを以下に示す。

- 1、 t についての積分区間 $[0, 1]$ を細かく分割し、各 t について 2、3 の手続きを行う。
- 2、 $\exp(-ntH(w))\varphi(w)$ に従うサンプル $\{w_k; k = 1, 2, \dots, K\}$ を取り出す。
- 3、

$$y(n, t) = \frac{1}{K} \sum_{k=1}^K nH(w_k)$$

を計算する。

- 4、 $\sum_t y(n, t)$ を目的の値とする。

この従来法について、上記のモデルを用いて実験を行った結果を図 4 に示す。ここで、実験の条件は実験 1 と同じようにした。この結果とアルゴリズム 2 での結果を比較すると、分散はほぼ同じであるが、誤差は大きい。また、表 1 より計算時間はほぼ同じとなった。計算精度に違いが出たのは、分割すること

でサンプル発生に用いる確率分布を単純にすることができ、その結果メトロポリス法で法則収束しやすくなったためであると考えられる。

5. 考 察

本論では、カルバック情報量 $H(w)$ と事前分布 $\varphi(w)$ で表される学習の順問題において、確率的複雑さが理論または実験で解明されている $H_1(w)$ があるときに、カルバック情報量を

$$H(w) = H_1(w) + H_2(w)$$

と分割することで確率的複雑さ $F(n)$ を計算する3方法を提案した。実験の結果、分割することで計算精度は向上し、また計算精度をある程度抑えた上で計算時間を大幅に短縮できることが確認できた。

本論で提案した方法に関連して、幾つかの観点から考察する。

(1) メトロポリス法では、 w から w' に更新される確率が 0.5 になるのがよいということが知られている [3]。そうなるためには用いる分布に応じた繰り返し回数とばらつきの設定が重要となる。今回の実験では、各アルゴリズムともに繰り返し回数とばらつきの条件を固定した上で比較を行ったが、最適な条件は各アルゴリズムで用いた分布が異なるためにそれぞれ違ったものになると思われる。よって、比較を行う際には条件を各アルゴリズムで最適化したうえで実験を行うのが望ましい。

しかしながら、最適化を行うにはカルバック情報量と事前分布が与えられたもとで、アルゴリズム 1、3 では各 n についてのみ行うのに対し、アルゴリズム 2 や分割しない方法では n だけでなく t についても最適化を行なう必要がある。

(2) 今回提案した方法で、学習係数 λ を求めることを考える。先に述べたように学習係数は確率的複雑さの主要項であり、このことから提案手法を用いることで学習係数を計算できる [5] [6]。そのためにはいろいろな n について確率的複雑さを計算する必要があり、特に n が十分大きくなったときの値が重要となる。そこで今回の実験で得られた性質より、 n があまり大きくないときには計算時間を少なくするためアルゴリズム 1 を用いて、 n が大きくなったときには分散を抑えるためアルゴリズム 3 を用いるとよいと思われる。

6. 結 論

カルバック情報量を分割して、確率的複雑さを計算するアルゴリズムを幾つか提案し、それらの性質と分割することの有効性を実験的に確認した。今後の課題としては、メトロポリス法の条件を最適化した上での提案アルゴリズムの性質の解明、提案手法を用いての学習係数の計算精度の確認などの問題が挙げられる。

この研究は、科学研究費補助金 15500130 の援助を受けた。

文 献

[1] 甘利俊一, 尾関智子, 朴慧暎, “階層的モデルにおける学習と推論-特異構造を持つ統計モデル”, 電子情報通信学会論文誌,

Vol.J85-DII, No.5, 701-708, 2002.
 [2] 青柳美輝, 渡辺澄夫, “特異点解消定理と学習理論への応用,” 信学技報, NC2-26, pp.25-30, 2003.
 [3] 伊庭幸人, “マルコフ連鎖モンテカルロ法とその統計学への応用”, 統計数理, 第 44 巻, 第 1 号, 49-84, 1996.
 [4] Yoshihiko Ogata, “A Monte Carlo Method for an objective Bayesian procedure,” Ann. Inst. Statist. Math., Vol.42, No.3, 403-433, 1990.
 [5] 永田賢二, 渡辺澄夫, “カルバック情報量の分割による特異モデルの学習係数の計算アルゴリズム” 情報論的学習理論 2003, 41-46, 2003.
 [6] 永田賢二, 渡辺澄夫, “カルバック情報量の分割による特異点モデルの学習係数計算法” 信学技報, NC2003-104, 67-72, 2003.
 [7] 西上功一郎, 渡辺澄夫, “特異な学習モデルの選択における事前分布の影響について” 電子通信情報学会論文誌, J86-D-II, No.1, 119-129, 2003.
 [8] 高橋克之, 渡辺澄夫, “特異的な学習モデルにおける MCMC 法の評価,” 情報論的学習理論, 2003.
 [9] 渡辺澄夫, “ベイズ法による階層型統計モデルの推定誤差について”, 電子情報通信学会論文誌, Vol.J81-A, No.10, 1442-1452, 1998.
 [10] 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, 16 巻 2 号, 308-315, 2001.
 [11] Sumio Watanabe, “Algebraic Analysis for Nonidentifiable Learning Machines”, Neural Computation, 13, 899-933, 2001.
 [12] 山崎啓介, 渡辺澄夫, “特異点をもつ推論モデルの学習曲線の確率的計算法”, 電子通信情報学会論文誌 D-II-85J(3), 363-372, 2002.