人工知能×特異点論＝？

日本大学理工学部数学科
青柳　美輝

1　はじめに

現在，社会は大きな変革を遂げ始めていると言われている．人工知能（AI）を用いて様々な場面で判断を行ったり，自宅の家電の操作をスマートホンなどからインターネットとの連携で可能とする IoT が利用されたりする Society5.0 とも呼ばれる新たな時代の到来が，社会や生活を大きく変えていくとの予測もある (図 1)．

本稿では学習理論の数理的な研究の立場から，人工知能に必要な機械学習に対する，代数幾何学とくに「特異点理論」の寄与について述べる．

計測して得られた多量のデータから，そのデータを発している情報源の確率分布（真の分布）を推測することを学習といい，学習の仕組みをまとめて体系化したものを「学習理論」と呼ぶ．パラメータを変化させることによって多くの確率分布を表現できる階層構造・内部構造をもつモデルは特異モデルと呼ばれ，複雑な構造を持っている．例えば画像音声認識・遺伝子解析・時系列予測・データマイニングに用いられる学習データは，正規分布に従うような単純なものではなく，その殆どが極めて複雑な構造を持つ．これらは，古典的理論の枠組みでは捉えることができないため，急速に多くの新理論の研究が始まった．とくに，学習効率を表す学習係数という概念においては，特異点解消定理との関係に着目する研究が重要となっている．

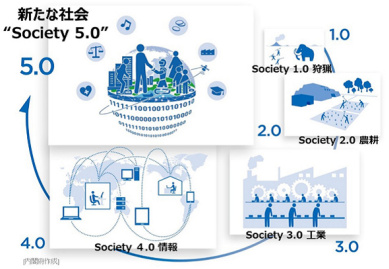
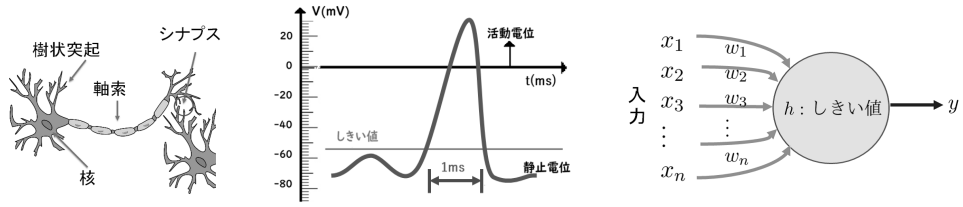


図 1: Society 5.0 : 狩猟社会 (Society 1.0), 農耕社会 (Society 2.0), 工業社会 (Society 3.0), 情報社会 (Society 4.0) に続く, 新たな社会 (Society 5.0). 出典：内閣府ホームページ (`https://www8.cao.go.jp/cstp/society5_0/index.html`)



(a) (b) (c)

図 2: (a) 生物の神経細胞 (b) 電気パルス (c) 神経細胞のモデル

2 ニューラルネットワーク

人工知能のモデルの一つとして，ニューラルネットワークについて紹介する．ニューラルネットワークは，生物の神経細胞の回路で行われている情報処理のモデル化である．神経細胞は，細胞体と呼ばれる本体，本体からつきだした樹状突起，他の細胞へつながる軸索からなる．軸索は，その末端が他の神経細胞の樹状突起に付着しており，軸索と他の神経細胞との結合部分をシナプスという．樹状突起で，他の細胞や感覚細胞から入力信号を受け，細胞体で信号処理をし，軸索，シナプスを通して，他の神経細胞に出力信号を出す (図 2 (a)).

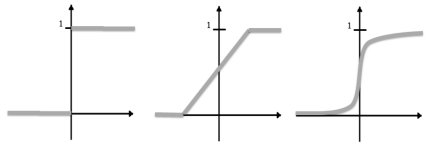
脳の神経細胞の数は 10 の 10 乗個から 10 の 11 乗個程度と言われている．これらの細胞を組み合わせて，各神経細胞が並列で分散して情報処理を行い，とても複雑で高度な処理が行われている．ひとつの細胞の出力は，10 から数百に分岐した軸索を通して，シナプスから他の細胞に伝えられる．ひとつの細胞が受けるシナプスの結合の数は，数百から数万に及ぶ．これらのすべてのシナプス結合が神経細胞間の信号の伝達に寄与している．

神経細胞では，他の神経細胞から信号が到着すると，その影響で膜電位が変化し，ある閾値を超えると電位が正の値に変化して興奮した状態になる．そして，他の神経細胞に信号を送る．このグラフの形は，入力された値に依存せず，ほぼ同一の波形であり，一度閾値を超えれば，形や大きさが一定の電気パルスが出る．したがって，ニューラルネットワークにおいて情報を担っているのは，電気パルスの波形ではなく，電気パルスの発生頻度であると考えられている (図 2 (b)).

閾値以上の入力があれば，電気パルスを出し，閾値以下であれば，出さないという，細胞体の閾値作用は，入力から出力への非線形変換効果をもつ．また，シナプスには，入力側の神経細胞を興奮しやすくする伝達物質を放出する興奮性シナプスと，逆に入力側の神経細胞を興奮しにくくさせる抑制性シナプスがある．受ける側の入力側神経細胞では，各出力側神経細胞からの入力の総和を受け取ると考えることができる．

ニューラルネットワークの数理モデルは，神経細胞の観察から生まれている．

1943 年，マッカロとピッツは，形式ニューロンモデルを提案した．図 2(c) の丸は，一つの神経細胞のモデルを表す．x_k は，0 と 1 の値を取り，この神経細胞が受け取るシナプ



(a) (b) (c)

図 3: (a) ステップ関数, (b) 区分線形関数, (c) シグモイド関数

ス前神経細胞の出力を表す．興奮した前神経細胞の出力 $x_k = 1$ が，この神経細胞に伝えられる．w_k はシナプス結合荷重，すなわちシナプス伝達効率とも呼ばれる．$w_k > 0$ のとき，興奮性シナプスに対応し，$w_k < 0$ のとき，抑制性シナプスに対応する．$|w_k|$ の値が大きければ，$w_k x_k$ として積で情報が伝達されるので，この神経細胞の電位に大きく影響を与える．

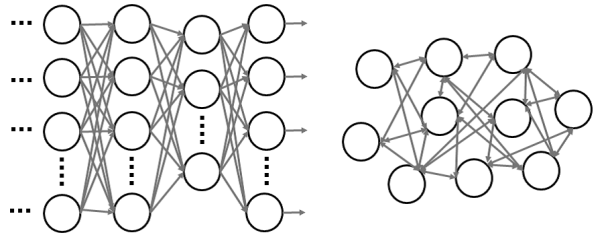
$\sum_{k=1}^{n} w_k x_k$ は，他の神経細胞からこの神経細胞が受ける電位変化の総和を表しており，閾値 h を超えると，この神経細胞は興奮し出力 $y = 1$ となり，次の細胞に情報を伝達する．閾値を超えない場合は，$y = 0$ である．$f(u) = \begin{cases} 1, & u > 0 \text{ のとき} \\ 0, & u \leq 0 \text{ のとき} \end{cases}$ とおけば，$y = f(\sum_{k=1}^{n} w_k x_k - h)$ と表される．

その後の研究においては，関数 f として，形式ニューロンモデルで用いられたステップ関数 (図 3(a)) 以外にも，連続な区分線形関数や微分可能な関数であるシグモイド関数 $f(u) = \dfrac{1}{1 + \exp(-\epsilon u)}$, $(\epsilon > 0)$ などが用いられている (図 3(b),(c))

ニューラルネットワークには，階層型ニューラルネットワークや相互結合型ニューラルネットワークなどのモデルがある (図 4). 階層型ニューラルネットワークは，心理学者ローゼンブラットにより提案されたモデルであり，入力層と出力層，中間層からなる．外部からの入力を入力層のニューロンが受け，出力層のニューロンが外部に出力を出す．相互結合型ニューラルネットワークは，物理学者のホップフィールドなどによって提案されたモデルであり，構成要素である各ニューロンは，他のすべてのニューロンと結合しており，各ニューロンの出力が平衡な値に収束することで情報処理が完了する．

階層型ニューラルネットワークの応用例として，画像認識がある．例として，ひらがなの文字認識の場合を考察する．適当な升目に区切られた文字に対して，一つ一つの小さな升目に文字の一部が入っている場合は，その升目に対応する入力層のニューロン値を 1 , そうでなければ，0 とする. 出力層には順番にそれぞれのひらがなに対応するニューロンを決定しておく．

図 5 は,「あ」という文字の認識の例である．この例では，20×20 個の升目に区切り，この小さな升目が入力層のニューロンに対応する．小さな枡の中に「あ」の文字の一部が入っている場合は 1 , そうでなければ 0 とする．すなわち，400 個の 0 と 1 のデータに変換する．図 5 の右端の「あ」の画像の升目が灰色のところに 1 という数値が入る．また，



(a) (b)

図 4: (a) 階層型ニューラルネットワーク，(b) 相互結合型ニューラルネットワーク

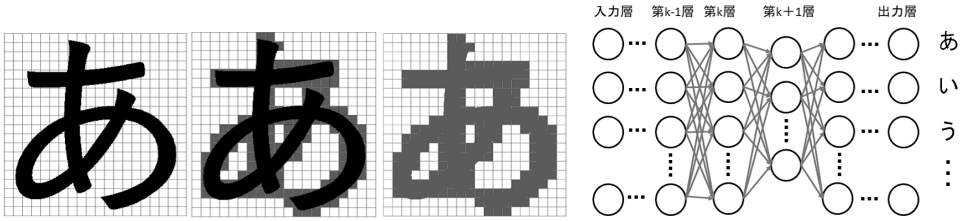


図 5: 文字認識の例

升目が白いところに 0 という数値が入る．その 400 個の 0 と 1 のデータを第一層の入力層に与える．最後の出力層では，1 番目を「あ」という文字に対応させ，2 番目を「い」という文字に対応させるというように順に決定する．

この機械を訓練する．ひらがなの文字を見せて,「あ」であれば，出力層の 1 番目のニューロンのみが出力 1 となるように，結合係数 w を適当な初期値から，あるアルゴリズムに従って変化させる．結合係数とは，ニューロン間の結びつきの強さを表している．画像を見せながら学習を繰りかえす．同様に,「い」などの他のひらがなについても訓練をする．それぞれのひらがなの訓練が終われば，この機械は，ひらがなの認識が可能になっている．これが機械学習の一つの例である．

3 学習理論

3.1 簡単な例

ここでは簡単な学習理論の例について考察する．x を入力，y を出力，w をパラメータとする．一般には，これらは多次元のベクトルである．学習モデルを $y = f(x, w)$ とする．階層型ニューラルネットワークでは x は入力層，y は出力層に対応し，w は結合係数である．真の関数 $y = h(x)$ から得られたデータを用いて，真の関数を学習する．すなわち最

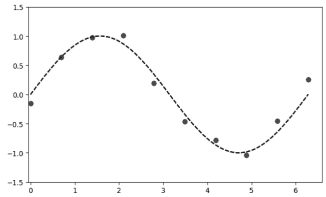


図 6: $y = \sin x +$ ノイズ　からのデータ 10 個

適なパラメータを見つけ，$y = f(x, w)$ が $y = h(x)$ になるべく近くなることを目指す．

この節では，評価関数として，最小二乗法を考える．

$\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$ を真の関数から得られたデータ $y_i = h(x_i) +$ ノイズ であるとする．Error $= \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i, w))^2$ が最小になるように w を決めることが最小二乗法である．

例として，$y = \sin x (0 \leq x \leq 2\pi)$ を真の関数として，$y = \sin x +$ ノイズ により 10 個のデータが得られるとする（図 6).

学習関数を $y = \sum_{i=0}^{m} w_i x^i$ とし，m 次の多項式で近似することを考える．ここで，x, y は一次元，$w = (w_0, w_1, \ldots, w_m)$ は $m + 1$ 次元である．$m = 2, 3, \ldots, 9$ 場合に Error $= \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i, w))^2$ が最小になるように w を決め，$\tilde{w}$ とする．図 7 は関数 $f(x, \tilde{w})$ を描いたものである．$m = 9$ の場合はすべてのデータの点を通っているが，$\sin x$ のグラフから離れてしまっている．

次に，汎化性を考察する．テストデータとして，$\{(x'_1, y'_1), (x'_2, y'_2), \cdots, (x'_{10}, y'_{10})\}$ を真の関数から新たに得られたデータとする．このデータに対して，$\tilde{w}$ を用いて $\frac{1}{n}\sum_{i=1}^{n}(y'_i - f(x'_i, \tilde{w}))^2$ をそれぞれ計算する．学習データの最小二乗誤差は小さくなっているが，テストデータの最小二乗誤差は $m = 3$ 以降大きくなっている．すなわち，このモデルでは $m = 3$ を選ぶのが妥当だと思われる．このように，実際にどのモデルを選ぶか，この場合は，m の値をどのように決定するかは，汎化性が重要なポイントである．

3.2　ベイズ法による学習理論

この節では，ベイズ法による学習理論について考察する．確率密度関数 $q(x, y)$ を真の分布，パラメータ w 付きの確率密度関数 $p(x, y|w)$ を学習モデルとする．

例えば，$y = h(x)+$「正規ノイズ」で得られるデータは，

$$q(x, y) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{|y - h(x)|^2}{2}) q(x)$$

から発生していると考えることができる．ここで，$q(x)$ は x の密度関数である．節 3.1 に

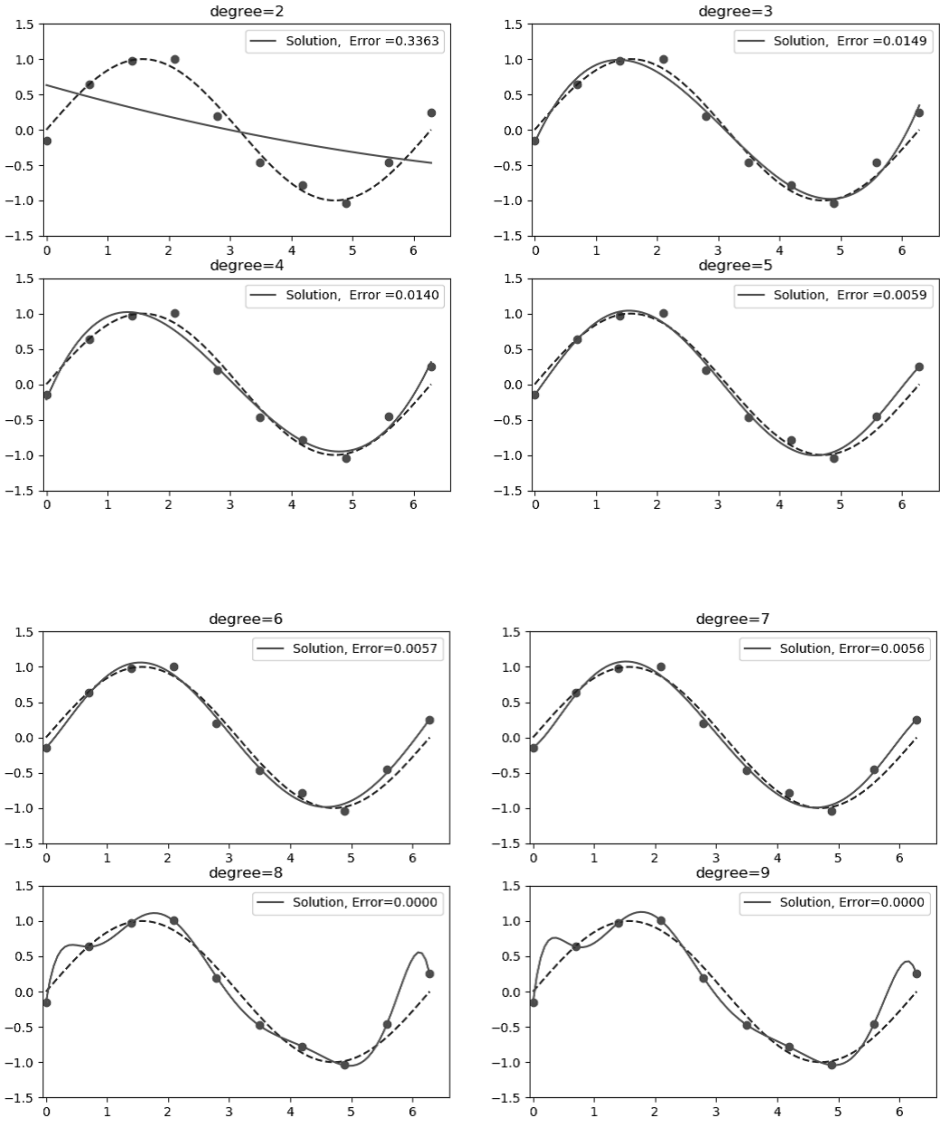


図 7: 最小二乗法による近似関数:degree $= 2 \sim 9$

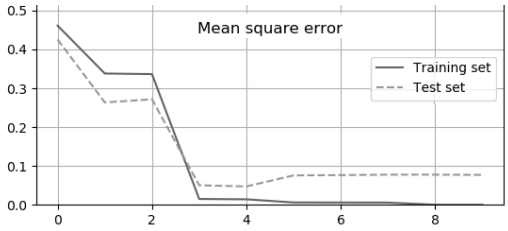


図 8: テストデータの最小二乗誤差 (点線）と，学習データの最小二乗誤差（実線）

おける例では，0 から 2π の一様分布とした．また，学習モデルは

$$p(x,y|w) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{|y-f(x,w)|^2}{2})q(x)$$

である．

確率密度関数 $q(x,y)$，$p(x,y|w)$ に対して，

$$K(w) = \int q(x,y) \log \frac{q(x,y)}{p(x,y|w)} dxdy$$

とおく．カルバック距離と呼ばれる相対エントロピーである．これは，$q(x,y)$ と $p(x,y|w)$ の「差」を表す．例えば，$q(x,y) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{|y-h(x)|^2}{2})q(x)$，
$p(x,y|w) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{|y-f(x,w)|^2}{2})q(x)$ の場合，y で積分することにより，

$$K(w) = \int q(x,y) \log \frac{q(x,y)}{p(x,y|w)} dydx = \frac{1}{2} \int (f(x,w)-h(x))^2 q(x)dx$$

となり，二乗誤差が現れる．

以降，(x,y) をまとめて，x と表す.

カルバック距離は，$K(w) \geq 0$ を満たし，$p = q$ のときのみ，$K(w) = 0$ という性質を持つ．証明は以下である．関数 $S(t) = \exp(-t) + t - 1, t \in \mathbf{R}$ は $S(t) \geq 0$ を満たす．また，

$$\begin{aligned}
\int S(\log \frac{q(x)}{p(x|w)}) q(x)dx &= \int \left(\exp\left(\log \frac{p(x|w)}{q(x)} \right) + \log \frac{q(x)}{p(x|w)} - 1 \right) q(x)dx \\
&= \int \left(\frac{p(x|w)}{q(x)} + \log \frac{q(x)}{p(x|w)} - 1 \right) q(x)dx \\
&= \int \left(q(x) \log \frac{q(x)}{p(x|w)} + p(x|w) - q(x) \right) dx \\
&= \int q(x) \log \frac{q(x)}{p(x|w)} dx + 1 - 1 = \int q(x) \log \frac{q(x)}{p(x|w)} dx.
\end{aligned}$$

したがって，任意の t に対して，$S(t) \geq 0$ であるから，$\int S(f(x,w))q(x)dx \geq 0$ である．また，等号が成り立つときは $p(x|w) = q(x)$ のときである．

統計学分野の学習理論における目的は，真の分布から発生する多量のデータセットから，そのデータを発している情報源の真の分布を再生・推測することである．汎化誤差は，推定された分布と真の分布とのエントロピーに関する誤差，カルバック距離を表している．従って，与えられたデータから求められる学習誤差から，汎化誤差を推定することは重要である．機械学習における文字認識，画像認識，音声認識，遺伝子解析などでは，データの情報源の確率分布は，正規分布に従うような単純なものではない．それらの機械学習においては，複雑な確率分布を表現できる階層構造・内部構造をもつニューラルネットワーク，混合正規分布や縮小ランクなどが利用されている．これらは，正規分布のような“よい性質”を持つ統計的正則モデルと区別するため，特異モデルと呼ばれている．古典的な理論の枠組みの中では捉えることができず，近年急速に多くの理論の研究が始まった．その中の一つとして，学習効率を表す学習係数の研究では，代数幾何の特異点解消（広中定理）との関係が指摘され，２つの異なる分野が融合し発展を始めている．

以下，それらの概要について説明する．

$x \in \mathbf{R}^N$ を確率変数，$q(x)$ を真の確率密度関数とし，$q(x)$ に従う n 個の独立なサンプルを $x^n := \{x_i\}_{i=1}^n$ とする．学習モデル $p(x|w)$ とその事前分布 $\psi(w)$ が与えられているものとする．ここで，パラメータ空間 $(w \in)W \subset \mathbf{R}^d$ はコンパクトとする．学習理論の目的は，データセット x^n から $p(x|w)$ を用いて，真の分布 $q(x)$ を推定することである．

ベイズ学習では，事後確率 $p(w|x^n)$ を

$$p(w|x^n) = \frac{1}{Z_n(\beta)}\psi(w)\prod_{i=1}^n p(x_i|w)^\beta,$$

で定義する．ここで Z_n は正規化定数

$$Z_n(\beta) = \int_W \psi(w)\prod_{i=1}^n p(x_i|w)^\beta \mathrm{d}w,$$

である．$\beta > 0$ は，逆温度とよばれ，通常 $\beta = 1$ である．確率密度関数 $p(w|x^n)$ は，得られたデータをもとにした，パラメータ w の確率である．学習モデル $p(x|w)$ は，w の値によって分布を変え，真の分布に近ければ，$p(x_i|w)$ は大きな値になり，$p(x|w)$ が真の分布からかけ離れていれば，$p(x_i|w)$ は 0 に近くなる．

ここで，

$$E_w^\beta[f(w)] = \int f(w)p(w|x^n)dw = \frac{\int dw f(w)\psi(w)\prod_{i=1}^n p(x_i|w)^\beta}{\int dw \psi(w)\prod_{i=1}^n p(x_i|w)^\beta},$$

および

$$V_w^\beta[f(w)] = E_w^\beta[f(w)^2] - E_w^\beta[f(w)]^2.$$

と定義する．この時，得られたデータ x^n から予測されたベイズ推測は $p(x|x^n) = E_w^\beta[p(x|w)]$ と定義される．

改めて，上記で述べたように確率密度関数 $p(x), q(x)$ に対して，カルバック距離 $K(q||p)$ を

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

また，経験カルバック距離 $K_n(q||p)$ を

$$K_n(q||p) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{q(x_i)}{p(x_i)},$$

で定義する．

ここで，汎化誤差 B_g, 学習誤差 B_t を次のように定義する．

$$B_g = K(q(x) \| p(x|x^n)),$$

$$B_t = K_n(q(x) \| p(x|x^n)).$$

汎化誤差は，真の分布を予測分布がどのくらい近似しているかを表したものである．また，汎化損失，経験損失をそれぞれ以下で定義する．

$$\begin{aligned} G_n &= -\int q(x) \log p(x|x^n) dx, \\ T_n &= -\frac{1}{n} \sum_{i=1}^{n} \log p(x_i|x^n). \end{aligned}$$

このとき，真の分布のみに依存する平均エントロピー $S = -\int q(x) \log q(x) dx$ および経験エントロピー $S_n = -\frac{1}{n} \sum_{i=1}^{n} \log q(x_i)$ に対して，

$$\begin{aligned} B_g &= G_n - S, \\ B_t &= T_n - S_n, \end{aligned}$$

が成り立つ．

$\lambda \in \mathbf{Q}$ を learning coefficient(学習係数)，$\nu \in \mathbf{R}$ を singular fluctuation とする．正規分布のような正則モデルでは，パラメータの次元を d とすると，$\lambda = \nu = d/2$ が成立する．

渡辺 [16] により以下の関係が示されている:

$$\begin{aligned} E[G_n] &= L(w_0) + \frac{1}{n} \left(\frac{\lambda - \nu}{\beta} + \nu \right) + o(\frac{1}{n}), \\ E[T_n] &= L(w_0) + \frac{1}{n} \left(\frac{\lambda - \nu}{\beta} - \nu \right) + o(\frac{1}{n}). \end{aligned}$$

ここで，$L(w) = -E_x[\log p(x|w)]$ および $w_0 \in W_0 = \{w \in W | L(w) = \min_{w' \in W} L(w')\}$ である．

学習係数 λ は，節 4 で詳しく述べるが，$E_x[\log(p(x|w_0)/p(x|w))]$ の log canonical threshold であり，θ をその位数とすると，それらの値を用いて，値 ν は，理論的に次で与えられる．

$$\nu = \frac{1}{2} E_\xi \frac{\int_0^\infty dt \sum_{u^*} \int du \xi(u) t^\lambda e^{-\beta t + \beta\sqrt{t}\xi(u)}}{\int_0^\infty dt \sum_{u^*} \int du t^{\lambda - 1/2} e^{-\beta t + \beta\sqrt{t}\xi(u)}}. \tag{1}$$

ここで，$\xi(u)$ は，特異点解消した空間上で定義された経験過程で，平均 0 分散 2 のガウス分布となる確率変数である．$\sum_{u^*}$ は，λ および θ が得られる局所座標の和である．

3.3 モデル選択

次にモデル選択について説明する．

3.3.1 sBIC および WBIC について

ベイズ推測においてモデルの選択を行う場合には，自由エネルギー

$$F_n(\beta) = -\frac{1}{\beta} \log \int \prod_{i=1}^n p(x_i|w)^\beta \psi(w) dw$$

を観測して比較する方法がある．

汎化誤差 と自由エネルギーには以下の関係がある．$n \geq 1, \beta = 1$ とする．このとき

$$\mathbb{E}[G_n] = \mathbb{E}[F_{n+1}(1)] - \mathbb{E}[F_n(1)], \quad \mathbb{E}[F_n(1)] = \sum_{i=1}^{n-1} \mathbb{E}[G_i] + \mathbb{E}[F_1(1)].$$

渡辺 [16] により

$$F_n(\beta) = nL_n(w_0) + \frac{\lambda}{\beta} \log(n) - \frac{\theta - 1}{\beta} \log\log(n) + O_p(1)$$

が成り立つ事が証明されている．ここで，$L_n(w) = -\frac{1}{n} \sum_{i=1}^n \log p(x_i|w), w_0 \in W_0$.

Drton[10] らは，$\beta = 1$ のときの，この関係式に対して「不動点方程式システム」により，近似計算する “singular Bayesian information criterion” (sBIC) 法を得た．これは学習係数 λ の理論値を用いる方法である．

一方，“widely applicable Bayesian information criterion” (WBIC) [17] 法は，理論値を用いず

$$WBIC = -E_w^\beta[\sum_{i=1}^n p(x_i|w)], \quad (\beta = 1/\log n),$$

と定義される．

モデル選択では，候補となるモデルに対して，これらの値を計算し，値の小さいモデルを選ぶ．この数値計算結果は最後の節 6 で紹介する．

3.3.2 WAIC およびクロスバリデーションについて

データ x^n に対して，$x^n \backslash x_i = \{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\}$ とおく．“widely applicable information criterion” (WAIC) [16] は

$$W_n = T_n + \frac{\beta}{n} \sum_{i=1}^{n} V_w^{\beta}[\log p(x_i|w)]$$

と定義され，クロスバリデーションは

$$C_n = -\frac{1}{n} \sum_{i=1}^{n} \log p(x_i|x^n \backslash x_i) \quad (n \geq 2)$$

と定義される．これらは以下の関係が示されている:

$$\begin{aligned} E[W_n] &= E[G_n] + o(\frac{1}{n}), \\ E[C_n] &= E[G_n] + o(\frac{1}{n}). \end{aligned}$$

前節 3.3.1 と同じように，モデル選択の場合，候補となるモデルに対して，これらの値を計算し，値の小さいモデルを選ぶ．これらは，汎化損失を，真の分布の情報を用いずに，観測データ x_i および学習モデル p を用いて推測できることを示している．

4 学習係数と log canonical threshold

学習係数は，代数幾何の分野では，カルバック関数の log canonical threshold として知られている．ここでは，近年得られた学習係数に関する結果とそれらを得るために必要な定理を紹介する．これらの理論値が求まると自由エネルギーと汎化誤差の理論値が明らかになるので，確率モデルの評価をするときに重要な役割を果たす．理論値は，数値計算の正しさを確認する手段ともなる．

定義 1 $\mathbb{C}^d$ または $\mathbb{R}^d$ における w^* の十分小さな近傍を U，f を U 上の 0 でない正則関数または実解析関数とする．$\psi(w)$ をコンパクトサポートを持つ C^{∞} 関数とする．このとき，f の w^* および ψ に関する log canonical threshold を，実数体上では，

$$c_{w^*}(f, \psi) = \sup\{c : \int_U |f|^{-c} \psi(w) dw < \infty\}$$

複素数体上では，

$$c_{w^*}(f,\psi) = \sup\{c : \int_U |f|^{-2c}\psi(w)dwd\overline{w} < \infty\}$$

と定義する．また，$\theta_{w^*}(f,\psi)$ をその位数とする．

$\psi(w^*) \neq 0$ ならば，特に，$c_{w^*}(f) = c_{w^*}(f,\psi)$ および $\theta_{w^*}(f) = \theta_{w^*}(f,\psi)$ とおく．これらの値は ψ に依存しないからである．

例 1 実数体上で，$f(x) = x^3$ であれば，$c_{w^*}(f) = 1/3, \theta_{w^*}(f) = 1$ である．なぜなら，$\int_0^1 x^{-3c}dx < \infty$ になるためには，$-3c+1 > 0$ であればよい．よって $c < 1/3$ である．

例 2 実数体上で，$f(x,y) = x^2 + y^2$ であれば，$c_{w^*}(f) = 1, \theta_{w^*}(f) = 1$ である．なぜなら，$x = r\cos\theta, y = r\sin\theta$ と極座標表示すれば，$\int_0^1 \int_0^{2\pi} r^{-2c}rdrd\theta < \infty$ となるので，$-2c+2 > 0$ であればよい．よって $c < 1$ である．

例 3 実数体上で，$f(x,y,z) = (xy)^2z$ であれば，$c_{w^*}(f) = 1/2, \theta_{w^*}(f) = 2$ である．なぜなら，$\int_0^1 \int_0^1 \int_0^1 x^{-2c}y^{-2c}z^{-c}dxdydz < \infty$ となるには，x, y の積分に注目して $-2c+1 > 0$ であればよい．よって $c < 1/2$ であり，位数は 2 である．

log canonical threshold は f に関するゼータ関数 $J(z) = \int_U |f|^z dw (z \in \mathbf{C})$ の最大の極でもある．代数幾何・代数解析では主に代数閉体上での log canonical threshold の研究が行われている．また，低次元での研究が主である ([12], [14])．例えば，複素体上での log canonical threshold は，代数解析における f の Bernstein-Sato 多項式 $b(s)$ の最大の根であることが知られている．

一方，学習理論における学習係数は，ある情報量の実数体上の log canonical threshold とその位数で与えられる．従ってそのまま複素体上の定理を学習係数に適用することができない．例えば，複素数体上の log canonical threshold は 1 より小さいが，実数体上ではそうとは限らない．学習理論における情報量の log canonical threshold は，ほとんどが 1 より大きい．ある関数族に関しては，実数体上の log canonical threshold の方が多くの情報を持っていることが知られている．このように学習係数を求めることは，数学的観点からも興味のある問題である．

log canonical threshold は広中の特異点解消定理により，原理的には有限の手続きにより求められるが，具体的に求めるのは難しいとされている．計算機で代数計算により行う方式も提案されているが，学習理論における特異点はパラメータを含んでいるため，確定された多項式の特異点解消よりも高度な面を含んでいる．更なる困難な問題点として，特異点が孤立していない・ニュートン図形が退化している等があげられる ([11], [18])．

今後 * を付加した記号 a^*, b^*, w^* などは定数を表すものとする．

補題 1 ([2, 3, 13]) U を $w^* \in \mathbb{R}^d$ の近傍，$\langle f_1, \cdots, f_n \rangle$ を U で定義された実解析関数 $f_1, \ldots, f_n$ で生成されるイデアルとする．

$g_1, \ldots, g_m \in \langle f_1, \cdots, f_n \rangle$ ならば $c_{w^*}(g_1^2 + \cdots + g_m^2) \leq c_{w^*}(f_1^2 + \cdots + f_n^2)$. 特に，$g_1, \ldots, g_m$ が $\langle f_1, \cdots, f_n \rangle$ の生成元ならば $c_{w^*}(f_1^2 + \cdots + f_n^2) = c_{w^*}(g_1^2 + \cdots + g_m^2)$.

定義 2 w^* の近傍 U で定義された実解析関数 $f_1, \cdots, f_m$ から生成されるイデアルを $\langle f_1, \cdots, f_n \rangle$ とする．このとき，$c_{w^*}(\langle f_1, \cdots, f_n \rangle) = c_{w^*}(f_1^2 + \cdots + f_m^2)$ とする．

この定義は，Lemma 1 より矛盾なく定義できる．

次の定理 2 は，複素数体において，関数を超平面に制限したときの log canonical threshold 値に関する定理である．

定理 2 ([12],[15]) $f(w_1, \cdots, w_d, w_{d+1})$ を原点の近傍における**正則関数**とする．g を $g = f|_{w_{d+1}=0}$ とおく．すなわち，f を $w_{d+1} = 0$ に制限した関数を g とする (または，H を超曲面として，f を H に制限した関数を $g = f_H$ とする).

このとき，$c_0(g) \leq c_0(f)$.

この定理は，実解析関数では成り立たない．たとえば，反例として，例 4 があげられる．

例 4

$$f(w_1, w_2, w_3, w_4) = (w_1^2 + w_2^2 + w_3^2 + w_4)^2,$$

とおく．このとき，$c_0(f) = 1/2$ であるが，$c_0(f(w_1, w_2, w_3, 0)) = 3/4$ となる．

しかし，斉次多項式の場合は，以下の様に成立する．

定理 3 ([4]) $f_1(w_1, \ldots, w_d)$, …, $f_m(w_1, \ldots, w_d)$ を 次数 n_i の $w_1, \cdots, w_d$ に対する斉次多項式とする．$f_1'(w_2, \ldots, w_d) = f_1(1, w_2, \ldots, w_d)$, …, $f_m'(w_2, \ldots, w_d) = f_m(1, w_2, \ldots, w_d)$ とおく．$w_1^* \neq 0$ であれば

$$c_{(w_1^*, \cdots, w_d^*)}(\langle f_1, \cdots, f_m \rangle) = c_{(w_2^*/w_1^*, \cdots, w_d^*/w_1^*)}(\langle f_1', \cdots, f_m' \rangle).$$

また，斉次多項式の場合は，以下の定理が成り立つ．

定理 4 ([4]) $f_1(w_1, \ldots, w_d)$, …, $f_m(w_1, \ldots, w_d)$ を $w_1, \cdots, w_j (j \leq d)$ に対する次数 n_i の斉次多項式とする．さらに，ψ を C^∞ 関数で，$\psi_{(0, \cdots, 0, w_{j+1}^*, \cdots, w_d^*)} \geq \psi_{(w_1^*, \cdots, w_d^*)}$ および $(0, \cdots, 0, w_{j+1}^*, \cdots, w_d^*)$ の近傍で $w_1, \cdots, w_j$ に関して，斉次であるとする．

このとき，

$$c_{(0, \cdots, 0, w_{j+1}^*, \cdots, w_d^*)}(\langle f_1, \cdots, f_m \rangle, \psi) \leq c_{(w_1^*, \cdots, w_j^*, w_{j+1}^*, \cdots, w_d^*)}(\langle f_1, \cdots, f_m \rangle, \psi)$$

が成り立つ．

一般に $w_0 \in U$ が

$$f_i(w_0) = \frac{\partial f_i}{\partial w_j}(w_0) = 0, 1 \leq i \leq m, 1 \leq j \leq d$$

を満たしたとしても

$$c_{w_0}(\langle f_1, \cdots, f_m \rangle, \psi) \leq c_{w^*}(\langle f_1, \cdots, f_m \rangle, \psi), w^* \in U$$

が成り立つとは限らない.

例 5 $f_1 = (x+1)x^2$, $f_2 = (y^2+x^2)((y-1)^6+x+1)$, $f_3 = (z^2+x^2)((z-1)^6+x+1)$ とする．このとき，$x = 0, y = 0, z = 0$ の時に限り，$f_1 = f_2 = f_3 = \frac{\partial f_1}{\partial x} = \frac{\partial f_2}{\partial y} = \frac{\partial f_2}{\partial x} = \frac{\partial f_3}{\partial z} = \frac{\partial f_3}{\partial x} = 0$ であるが，$c_{(0,0,0)}(\langle f_1, f_2, f_3 \rangle) = 3/4 > c_{(-1,1,1)}(\langle f_1, f_2, f_3 \rangle) = 2/3$.

これらの定理は，実数体上の log canonical threshold である学習係数を求めるために重要な役割を果たす．log canonical threshold を得る最良の特異点を得ることや，変数の追加により blow-up プロセス内で行われる変数変換の簡略化，blow-up プロセス数を減少させるのに有効である.

5 これまで得られている学習係数の結果

ここでは，学習理論においてよく用いられる混合正規分布，三層ニューラルネットワーク，混合二項分布のベイズ推測に関する学習効率を与える Vandermonde matrix singularities の場合 [1, 4, 5, 6, 7, 8] と，学習係数がすべての場合に明らかになっている縮小ランクモデルの場合 [9] における log canonical threshold に関する結果を述べる.

5.1 縮小ランクモデルの log canonical threshold

縮小ランクモデルは，学習係数 λ の理論値が以下のようにすべて解明されている [9].

パラメータの集合を $\{w = (A, B) \mid A \in N \times H$ 行列, $B \in H \times M$ 行列 $\}$ とする．入力 $x \in \mathbf{R}^M$ を密度関数 $q(x)$ から生成されるものとし，出力を $y = ABx +$ 正規ノイズ $\in \mathbf{R}^N$ とする．学習モデルは,

$$p(x, y|w) = \frac{1}{(\sqrt{2\pi})^N} \exp(-\frac{1}{2}||y - ABx||^2)q(x)$$

となる.

定理 5 真のパラメータを $w=(A_0,B_0)$. r を A_0B_0 のランクとするとき，λ，θ は次で与えられる:

(1) $N+r\leq M+H$, $M+r\leq N+H$, $H+r\leq M+N$.

(a) $M+H+N+r$ が偶数ならば，　$\theta=1$,

$$\lambda=\frac{-(H+r)^2-M^2-N^2+2(H+r)M+2(H+r)N+2MN}{8}.$$

(b) $M+H+N+r$ が奇数ならば，$\theta=2$,

$$\lambda=\frac{-(H+r)^2-M^2-N^2+2(H+r)M+2(H+r)N+2MN+1}{8}.$$

(2) $M+H<N+r$ ならば，$\theta=1$，$\lambda=\dfrac{HM-Hr+Nr}{2}$.

(3) $N+H<M+r$ ならば $\theta=1$，$\lambda=\dfrac{HM-Hr+Nr}{2}$.

(4) $M+N<H+r$ ならば $\theta=1$，$\lambda=\dfrac{MN}{2}$.

5.2 Vandermonde matrix 型特異点の log canonical threshold

定義 3 $Q\in\mathbf{N}$ を固定する.

$b_1^*=\cdots=b_{i-1}^*=0$, $b_i^*\neq 0$ のとき，$\gamma_i=\begin{cases}1, & Q\text{ が奇数}\\ \mathrm{sign}(b_i^*)\text{ (符号)}, & Q\text{ が偶数}\end{cases}$ に対して，$[b_1^*,b_2^*,\cdots,b_N^*]_Q=\gamma_i(0,\cdots,0,b_i^*,\cdots,b_N^*)$ と定義する.

定義 4 $Q\in\mathbf{N}$ を固定する.

$$A=\begin{pmatrix} a_{11} & \cdots & a_{1H} & a_{1,H+1}^* & \cdots & a_{1,H+r}^*\\ a_{21} & \cdots & a_{2H} & a_{2,H+1}^* & \cdots & a_{2,H+r}^*\\ \vdots & & & \vdots & & \\ a_{M1} & \cdots & a_{MH} & a_{M,H+1}^* & \cdots & a_{M,H+r}^* \end{pmatrix},\ I=(\ell_1,\ldots,\ell_N)\in {\mathbf{N}_{+0}}^N$$

$$B_I=(\prod_{j=1}^N b_{1j}^{\ell_j},\prod_{j=1}^N b_{2j}^{\ell_j},\cdots,\prod_{j=1}^N b_{Hj}^{\ell_j},\prod_{j=1}^N {b_{H+1,j}^*}^{\ell_j},\cdots,\prod_{j=1}^N {b_{H+r,j}^*}^{\ell_j})^t$$

$$B=(B_I)_{\ell_1+\cdots+\ell_N=Qn+1,0\leq n\leq H+r-1}$$

$$=(B_{(1,0,\cdots,0)},B_{(0,1,\cdots,0)},\cdots,B_{(0,0,\cdots,1)},B_{(1+Q,0,\cdots,0)},\cdots)$$

とする (t は行列の転置を表す).

a_{ki}，b_{ij} $(1\leq k\leq M,1\leq i\leq H,1\leq j\leq N)$ は，定数 a_{ki}^*，b_{ij}^* の近傍で定義された変数とする.

J を AB のすべての成分から生成されるイデアルとする.

J で定められる特異点を Vandermonde matrix 型特異点とよぶ. 簡単のため, $1 \leq j \leq r$ に対して,

$$(a^*_{1,H+j}, a^*_{2,H+j}, \cdots, a^*_{M,H+j})^t \neq 0, (b^*_{H+j,1}, b^*_{H+j,2}, \cdots, b^*_{H+j,N}) \neq 0$$

および $j \neq j'$ に対して,

$$[b^*_{H+j,1}, b^*_{H+j,2}, \cdots, b^*_{H+j,N}]_Q \neq [b^*_{H+j',1}, b^*_{H+j',2}, \cdots, b^*_{H+j',N}]_Q$$

を仮定する.

例 6 $N = r = 1$, $M = Q = 2$, $H = 3$ の場合,

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a^*_{14} \\ a_{21} & a_{22} & a_{23} & a^*_{24} \end{pmatrix}, B = \begin{pmatrix} b_{11} & b^3_{11} & b^5_{11} & b^7_{11} \\ b_{21} & b^3_{21} & b^5_{11} & b^7_{21} \\ b_{31} & b^3_{31} & b^5_{11} & b^7_{31} \\ b^*_{41} & {b^*_{41}}^3 & {b^*_{41}}^5 & {b^*_{41}}^7 \end{pmatrix}$$

となる. $Q = 2$ の場合は, 入力層ニューロンの個数 M, 中間層ニューロンの個数 H, 出力層ニューロンの個数 N の三層ニューラルネットワークに対応する.

例 7 $Q = M = r = 1$, $H = N = 2$ の場合,

$$A = \begin{pmatrix} a_{11} & a_{12} & a^*_{13} \end{pmatrix}, B = \begin{pmatrix} b_{11} & b_{12} & b^2_{11} & b_{11}b_{12} & b^2_{12} & b^3_{11} & b_{11}b^2_{12} & b^2_{11}b_{12} & b^3_{12} \\ b_{21} & b_{22} & b^2_{21} & b_{21}b_{22} & b^2_{22} & b^3_{21} & b_{21}b^2_{22} & b^2_{21}b_{22} & b^3_{22} \\ b^*_{31} & b^*_{32} & {b^*_{31}}^2 & b^*_{31}b^*_{32} & {b^*_{32}}^2 & {b^*_{31}}^3 & b^*_{31}{b^*_{32}}^2 & {b^*_{31}}^2 b^*_{32} & {b^*_{32}}^3 \end{pmatrix}.$$

$Q = 1$, $M = 1$ の場合は, 峰の個数 H の混合正規分布に対応する.

以下, 次のように定義する: $A_{M,H} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1H} \\ a_{21} & a_{22} & \cdots & a_{2H} \\ & & \vdots & \\ a_{M1} & a_{M2} & \cdots & a_{MH} \end{pmatrix}$,

$$B_{H,N,I} = \begin{pmatrix} \prod_{j=1}^N {b_{1j}}^{\ell_j} \\ \prod_{j=1}^N {b_{2j}}^{\ell_j} \\ \vdots \\ \prod_{j=1}^N {b_{Hj}}^{\ell_j} \end{pmatrix}, B^{(Q)}_{H,N} = (B_{H,N,I})_{\ell_1+\ldots+\ell_N = Qn+1, 0 \leq n \leq H-1}.$$

さらに $\mathbf{a}^* = \begin{pmatrix} a^*_{1,H+1} \\ \vdots \\ a^*_{M,H+1} \end{pmatrix}$ および $(A_{M,H}, \mathbf{a}^*) = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1H} & a^*_{1,H+1} \\ a_{21} & a_{22} & \cdots & a_{2H} & a^*_{2,H+1} \\ & & \vdots & & \\ a_{M1} & a_{M2} & \cdots & a_{MH} & a^*_{M,H+1} \end{pmatrix}$

とおく.

次の定理は $c_0(||A_{M,H}B^{(Q)}_{H,N}||^2)$ および $c_0(||(A_{M,H-1},\mathbf{a}^*)B^{(Q)}_{H,N}||^2)$ の値がわかれば，すべての Vandermonde matrix 型特異点の log canonical threshold がわかることを示している．

定理 6 ([3]) U を

$$w^* = \{a^*_{ki}, b^*_{ij}\}_{1\leqq k\leqq M, 1\leq i\leq H, 1\leq j\leq N}$$

の近傍とし，変数 $w = \{a_{ki}, b_{ij}\}_{1\leqq k\leqq M, 1\leq i\leq H, 1\leq j\leq N}$ は U 内に値を取るとする．

$(b^{**}_{01}, b^{**}_{02}, \cdots, b^{**}_{0N}) = (0, \ldots, 0)$ とおく．

ここで

$$[b^*_{i1}, b^*_{i2}, \cdots, b^*_{iN}]_Q \neq 0, \quad i = 1, \ldots, H+r$$

の中で，異なるベクトルを $(b^{**}_{11}, b^{**}_{12}, \cdots, b^{**}_{1N})$, …, $(b^{**}_{r'1}, b^{**}_{r'2}, \cdots, b^{**}_{r'N})$ とする．すなわち，

$$\{(b^{**}_{11}, \cdots, b^{**}_{1N}), \ldots, (b^{**}_{r'1}, \cdots, b^{**}_{r'N})\ ;\ [b^*_{i1}, \cdots, b^*_{iN}]_Q \neq 0, i = 1, \ldots, H+r\}.$$

このとき r' は，一意的に定まり，仮定より $r' \geq r$ である．

$1 \leq i \leq r$ に対して，$(b^{**}_{i1}, \cdots, b^{**}_{iN}) = [b^*_{H+i,1}, \cdots, b^*_{H+i,N}]_Q$, とする．

$$[b^*_{i1}, \cdots, b^*_{iN}]_Q = \begin{cases} 0, & 1 \leq i \leq H_0 \\ (b^{**}_{11}, \cdots, b^{**}_{1N}), & H_0+1 \leq i \leq H_0+H_1, \\ (b^{**}_{21}, \cdots, b^{**}_{2N}), & H_0+H_1+1 \leq i \leq H_0+H_1+H_2, \\ \vdots & \\ (b^{**}_{r'1}, \cdots, b^{**}_{r'N}), & H_0+\cdots+H_{r'-1}+1 \leq i \leq H_0+\cdots+H_{r'}, \end{cases}$$

および $H_0 + \cdots + H_{r'} = H$ としておく．

このとき

$$c_{w^*}(||AB||^2) = \frac{Mr'}{2} + c_{w_1^{(0)*}}(||A_{M,H_0}B^{(Q)}_{H_0,N}||^2)$$

$$+ \sum_{\alpha=1}^{r} c_{w_1^{(\alpha)*}}(||(A_{M,H_\alpha-1}, \mathbf{a}^{(\alpha)*})B^{(1)}_{H_\alpha,N}||^2) + \sum_{\alpha=r+1}^{r'} c_{w_1^{(\alpha)*}}(||A_{M,H_\alpha-1}B^{(1)}_{H_\alpha-1,N}||^2).$$

ここで，$w_1^{(0)*} = \{a^*_{k,i}, 0\}_{1\leq i\leq H_0}$,

$w_1^{(\alpha)*} = \{a^*_{k,H_0+\cdots+H_{\alpha-1}+i}, 0\}_{2\leq i\leq H_\alpha}$ および $\mathbf{a}^{(\alpha)*} = \begin{pmatrix} a^*_{1,H+\alpha} \\ \vdots \\ a^*_{M,H+\alpha} \end{pmatrix}$ $(\alpha \geq 1)$.

$N = 1$ の場合は, 次のような結果が得られている.

定理 7 ([2]) $N = 1$ のとき $\lambda = \lambda' = \frac{MQk(k+1)+2H}{4(1+kQ)}$. ここで，$k = \max\{i \in \mathbb{Z}; 2H \geq M(i(i-1)Q+2i)\}$ である．

$$\theta = \begin{cases} 1, & if\ 2H > M(k(k-1)Q+2k) \\ 2, & if\ 2H = M(k(k-1)Q+2k) \end{cases}$$

また,

$$\theta' = \begin{cases} 1, & if\ M = H = 1 \\ 1, & if\ 2H > M(k(k-1)Q + 2k) \\ 2, & if\ 2H = M(k(k-1)Q + 2k), H > 1 \end{cases}$$

次の定理は，学習係数の上界を与える.

定理 8 ([4])

bound_1

$$= \min\{\frac{(H-i+1)N + d_i(s) + d_i'(s) + d_i''(s)}{2(c(i,s,k(s))-1)Q+2} : 1 \leq i \leq s, 1 \leq k(1), \cdots, k(s) \leq N, 1 \leq s \leq H\}$$

ここで
$c(i,s,j) = \#\{i_1 : i \leq i_1 \leq s, k(i_1) = j\}$, $C(i,s) = \#\{c(i,s,j) = 0, 1 \leq j \leq N\}$,

$$\begin{aligned} d_i(s) &= (N-1)Q \sum_{s_1=i}^{s} (c(i,s_1,k(s_1)) - 1) \\ d_i'(s) &= M(i-1)\{(c(i,s,k(s))-1)Q + 1\} \\ &\quad + QM \sum_{\substack{s_1=i, \\ c(i,s,k(s)) > c(i,s_1,k(s_1))}}^{s-1} (c(i,s,k(s)) - c(i,s_1,k(s_1))), \end{aligned}$$

$$d_i''(s) = \begin{cases} 0, & if\ c(i,s,k(s)) = 1, \\ (H-s)\{C(i,s)Q + (N-1)Q(c(i,s,k(s))-2)\}, & \\ \quad if\ c(i,s,k(s)) \geq 2, N-1 \leq M, & \\ (H-s)\{C(i,s)Q + MQ(c(i,s,k(s))-2)\}, & \\ \quad if\ c(i,s,k(s)) \geq 2, C(i,s) \leq M < N-1, & \\ (H-s)\{MQ(c(i,s,k(s))-1)\}, & \\ \quad if\ c(i,s,k(s)) \geq 2, M \leq C(i,s). & \end{cases}$$

とする.

さらに，$\langle {k \atop l} \rangle = \frac{k!}{l!(k-l)!}$ に対して，$\mathrm{bound}_2 = \frac{NH + \sum_{i=0}^{k'-1} MQ(k'-i)\langle {N+Qi \atop N-1} \rangle}{2+2Qk'}$ とする.
ここで $k' = \max\{i \in \mathbb{Z}; NH \geq M\sum_{i'=0}^{i-1}(1+Qi')\langle {N+Qi' \atop N-1} \rangle\}$ である.
また

$$\mathrm{bound}_3 = \frac{MH}{2}$$

とする.
このとき

$$c_0(||A_{M,H}B_{H,N}^{(Q)}||^2) \leq \min\{\mathrm{bound}_1, \mathrm{bound}_2, \mathrm{bound}_3\}$$
$$c_0(||(A_{M,H-1}, \mathbf{a}^*)B_{H,N}^{(Q)}||^2) \leq \min\{\mathrm{bound}_1, \mathrm{bound}_2\}$$

次に，$H = 1, 2, 3$ の場合の真の値を与える．$\lambda = c_0(\|A_{M,H}B_{H,N}^{(Q)}\|^2)$ とし，θ をその位数とする．また，$\lambda' = c_0(\|(A_{M,H-1}, \mathbf{a}^*)B_{H,N}^{(Q)}\|^2)$, および θ' をその位数とする.

定理 9 Case 1 $H = 1$.

1. $\lambda = \min\{\frac{M}{2}, \frac{N}{2}\}, \theta = \begin{cases} 1, & \text{if } M \neq N, \\ 2, & \text{if } M = N, \end{cases}$

2. $\lambda' = \frac{N}{2}, \theta' = 1.$

Case 2 $H = 2.$

1. $M > N+1$ ならば $\lambda = \lambda' = N,\ \theta = \theta' = 1.$

2. $M = N+1$ ならば $\lambda = \lambda' = N,\ \theta = \theta' = 2.$

3. $M = N$ ならば $\lambda = \lambda' = \frac{2N+Q(2N-1)}{2(Q+1)},\ \theta = \theta' = 1.$

4. $M \leq N-1$ ならば $\lambda = M,\ \theta = 1.$

5. $N-Q+1 \leqq M \leq N-1$ ならば $\lambda' = \frac{2N+Q(2N-1)}{2(Q+1)},\ \theta' = 1.$

6. $M = N-Q$ ならば $\lambda' = \frac{N+M}{2},\ \theta' = 2.$

7. $M \leq N-Q-1$ ならば $\lambda' = \frac{N+M}{2},\quad \theta' = 1.$

Case 3 $H = 3.$

1. $M > N+2$ ならば $\lambda = \lambda' = \frac{3N}{2},\ \theta = \theta' = 1.$

2. $M = N+2$ ならば $\lambda = \lambda' = \frac{3N}{2},\ \theta = \theta' = 2.$

3. $M = N+1$ ならば $\lambda = \lambda' = \frac{3N+(3N-1)Q}{2(Q+1)},\ \theta = \theta' = 1.$

4. $M = N$ ならば $\lambda = \lambda' = \frac{3N+(3N-2)Q}{2(Q+1)},\ \theta = \theta' = 2.$

5. $M = N-1$ ならば $\begin{cases} \lambda = \frac{3-Q+3M(Q+1)}{2(Q+1)},\ \theta = 1\ (Q > 3), \\ \lambda = \frac{3M}{2},\ \theta = 2\ (Q = 3), \\ \lambda = \frac{3M}{2},\ \theta = 1\ (Q < 3). \end{cases}$

6. $M < N-1$ ならば $\lambda = \frac{3M}{2},\quad \theta = 1.$

7. $M = N-S\ (S = 1, 2, \cdots)$ ならば
$\begin{cases} \lambda' = \frac{S(3+Q)-2Q+3M(Q+1)}{2(Q+1)},\ \theta' = 1\ (Q > S), \\ \lambda' = \frac{2M+N}{2},\ \theta' = 2\ (Q = S), \\ \lambda' = \frac{2M+N}{2},\ \theta' = 1\ (Q < S). \end{cases}$

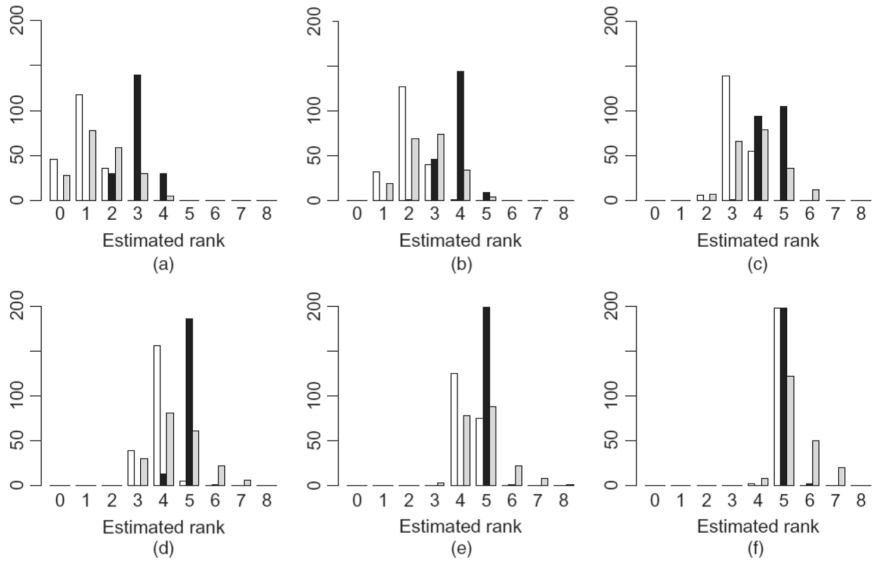


図 9: 論文 [10] から引用：Frequencies of rank estimates in reduced rank regression by using Schwarz's BIC (□), WBIC (▒) and sBIC (■) (results from 200 simulations with 10× 15 matrices of true rank 5): (a) $n = 50$; (b) $n = 100$; (c) $n = 200$; (d) $n = 300$; (e) $n = 500$; (f) $n = 1000$

論文 [5] では，$M = 4$ の場合まで得られている．

Vandermonde matrix 型特異点は，今日の実世界データに比較的頻繁に応用される，混合正規分布，三層ニューラルネットワーク，混合二項分布の学習係数を与えるものである．異なるモデルから共通の特異性が現れたことから，学習理論において本質的ではないかと考えられている．また，縮小ランクそのものも応用範囲が広いが，近年，ニューラルネットワークではシグモイド関数ではなく区分線形関数を用いる事も多い．したがって，その結果は，そのままニューラルネットワークに応用できる可能性が高い．

6 数値計算

論文 [10] から縮小ランクモデルに対する数値計算を引用する．数値実験では，$N = 10$, $M = 15$ とし，適当なランク 5 の真の分布を 200 個用意する．真の分布から，n 個のデータをランダムに生成する．実験では，正則モデルの場合に利用される Schwarz's BIC 法と，節 3.3.1 で紹介した Drton らによる sBIC 法，および，渡辺による WBIC 法を比較している (図 9). 理論値 λ を利用した sBIC は比較的 n の値が小さくても，正しくモデル選択を行っていることが分かる．また，sBIC は定理 8 で求めたような上界でも利用できることが証明されている．

7　最後に

$\beta = 1$ のとき，汎化損失は，$E[G_n] = L(w_0) + \frac{\lambda}{n} + o(\frac{1}{n})$，自由エネルギーは，$F_n(1) = nL_n(w_0) + \lambda \log(n) - (\theta - 1) \log\log(n) + O_p(1)$ となる．これらは，学習理論において代表的な指標であり，大変重要な値である．これらの主要項は，代数幾何などで定義された log canonical threshold $(= \lambda)$ から得られ，この理論値が求まれば，汎化損失や自由エネルギーの挙動を知ることができる．真の分布が分からないという状況においては，これらの理論値は解析のための重要な礎となる．また，理論値が分かっていれば，確率モデルの評価はもちろん，事後分布を数値的に実現したときに，その実現アルゴリズムが事後分布をよく近似しているかどうかを確認することができる．このように，理論値は，数値計算の正しさを確認する手段ともなる．

このような特異点論からの考察はベイズ推測と他の推測法の精度の違いを明らかにすることができるため，特異モデルの場合，最尤推測，事後確率最大化推測は適切ではなく，ベイズ推測が適していることも証明されている．

人工知能は，現在，応用面では急速に発展している．しかし，実験的経験的な観点から議論されることは多いが，理論的な解析はまだ多くの部分で進んでいないように思われる．したがって，特異点論を取り入れた学習理論を用いて解析をしていく方法は，これから非常に重要な役割を果たすのではないだろうか．

新型コロナウィルスの影響で，講演はなくなりましたが，このように原稿を書く機会を与えていただき大変感謝しております．

参考文献

[1] M. Aoyagi. The zeta function of learning theory and generalization error of three layered neural perceptron. *RIMS Kokyuroku, Recent Topics on Real and Complex Singularities*, 1501:153–167, 2006.

[2] M. Aoyagi. Log canonical threshold of Vandermonde matrix type singularities and generalization error of a three layered neural network. *International Journal of Pure and Applied Mathematics*, 52(2):177–204, 2009.

[3] M. Aoyagi. A Bayesian learning coefficient of generalization error and Vandermonde matrix-type singularities. *Communications in Statistics - Theory and Methods*, 39(15):2667–2687, 2010.

[4] M. Aoyagi. Consideration on singularities in learning theory and the learning coefficient. *Entropy*, 15(9):3714–3733, 2013.

[5] M. Aoyagi. Learning coefficient of vandermonde matrix-type singularities in model selection. *Entropy (Information Theory, Probability and Statistics)*, 21(6-561):1–12, 2019.

[6] M. Aoyagi. Learning coefficients and information criteria. *Frontiers in Artificial Intelligence and Applications*, pages 351–362, 2019.

[7] M. Aoyagi and K. Nagata. Learning coefficient of generalization error in Bayesian estimation and Vandermonde matrix type singularity. *Neural Computation*, 24(6):1569–1610, 2012.

[8] M. Aoyagi and S. Watanabe. Resolution of singularities and the generalization error with Bayesian estimation for layered neural network. *IEICE Trans. J88-D-II*, 10:2112–2124, 2005a.

[9] M. Aoyagi and S. Watanabe. Stochastic complexities of reduced rank regression in Bayesian estimation. *Neural Networks*, 18:924–933, 2005b.

[10] M. Drton and M. Plummer. A bayesian information criterion for singular models. *J. R. Statist. Soc. B*, 79(2):1–38, 2017.

[11] H. Hironaka. Resolution of singularities of an algebraic variety over a field of characteristic zero. *Annals of Math*, 79:109–326, 1964.

[12] J. Kollár. Singularities of pairs. *Algebraic geometry-Santa Cruz 1995, Proc. Symp. Pure Math., American Mathematical Society, Providence, RI,*, 62:221–287, 1997.

[13] S. Lin. Asymptotic approximation of marginal likelihood integrals. *(preprint)*, 2010.

[14] M. Mustata. Singularities of pairs via jet schemes. *J. Amer. Math. Soc.*, 15:599–615, 2002.

[15] T. Ohsawa and K. Takegoshi. On the extension of L^2 holomorphic functions. *Math. Zeitschrift*, 195:197–204, 1987.

[16] S. Watanabe. *Algebraic Geometry and Statistical Learning Theory*, volume 25. Cambridge University Press, New York, USA, 2009.

[17] S. Watanabe. A widely applicable bayesian information criterion. *Journal of Machine Learning Research*, (14):867–897, 2013.

[18] S. Watanabe, K. Hagiwara, S. Akaho, Y. Motomura, K. Fukumizu, M. Okada, and M. Aoyagi. *Theory and Application of Learning System.* Morikita Press, 2005.