CASを用いた損失関数の応用
- Application of loss functions using CAS -

奈良工業高等専門学校 一般教科　鷲野朋広

Tomohiro Washino

Department of Liberal Studies, National Institute of Technology (KOSEN), Nara College

羽衣国際大学 現代社会学部 高橋正

Tadashi Takahashi

Faculty of Social Sciences, Hagoromo University of International Studies

Abstract

The statistical model is a three-layer neural network with two intermediate units. Near the singular region where the number of intermediate units changes from 2 to 1, a plateau phenomenon occurs where learning stagnates[1],[2]. Amari et al. fixed fast-changing parameters and examined the dynamics of slow-changing parameters [3]. Guo et al. examined five patterns of learning dynamics in the near singular regions [4]. It is known that the set of parameters that realize the true distribution by Watanabe can be expressed as a set of zeros of finite polynomials by series expansion of the activation function [5],[6]. We transform the coordinate system of the polynomial and use Mathematica to obtain the parameter set and singularity set of the statistical model when the number of hidden units is 2 and the true distribution is realized with hidden units of 1.

1 Introduction

Definition 1 (True density function, statistical model)
The n independent random variables $X^n = (X_1, X_2, ..., X_n)$ following a probability density function $q(x)$ on $\mathbb{R}$ are called examples. The probability distribution $q(x)$ for which the example holds is called the true distribution. The probability density function $p(x|\theta)$ of $\theta \in W \subset \mathbb{R}^d$ is called the statistical model.

Definition 2 (Set of true parameters)
For $W \subset \mathbb{R}^d$, $q(x)$, $p(x|\theta)$, the true parameter set W_0 is defined as follows:

$$W_0 := \{\theta \in W \,|\, For\, all\, x,\, q(x) = p(x|\theta)\}\,.$$

Definition 3 (Loss Functions)
To estimate the true distribution $q(x)$, we determine the empirical log-loss function $L_n(\theta)$ as follows:

$$L_n(\theta) := -\frac{1}{n}\sum_{i=1}^{n} \log p(X_i|\theta).$$

When $\theta \in W_0$ for, $S_n := -\frac{1}{n}\sum_{i=1}^{n} \log q(X_i)$, let $f(x,\theta) := \log \frac{q(x)}{p(x|\theta)}$ be the log likelihood ratio function, it satisfies as follows : $L_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} f(X_i,\theta) + S_n$.

Find the parameter $\hat{\theta}$ that minimizes the empirical loss $\frac{1}{n}\sum_{i=1}^{n} f(X_i, \theta)$, The method in which the probability density function $p(x|\hat{\theta})$ is used as the guess result is called the maximum likelihood estimation method.

Definition 4 (Kullback information)
For the log-likelihood ratio function $f(x, \theta)$, the generalization loss (Kullback information) $K(\theta)$ is obtained as follows:

$$K(\theta) := \int q(x) f(x, \theta) dx.$$

In particular, the set of true parameters is the set of parameters for which the Kullback information is zero, and the following holds:

$$W_0 := \left\{ \theta \in W \,\middle|\, \int q\left(x\right) \log \frac{q\left(x\right)}{p\left(x|\theta\right)} dx = 0 \right\}.$$

Definition 5
When the input X and output Y follow the distribution $q(x, y) = q(y|x)q(x)$, we define the statistical model as a conditional probability density function $p(y|x, \theta)$.

Definition 6
For a probability distribution $q(x)$ followed by an input X and a function $f(x, \theta)$ from $\mathbb{R}$ to $\mathbb{R}$ with parameter θ , consider a random variable Z on $\mathbb{R}$ with mean 0 and variance 1, A random variable Y on $\mathbb{R}$ is called a function approximation model is as follows; $Y := f(x, \theta) + Z$.

For $\theta = (w_1,\, w_2,\, w_{31},\, w_{32}) \in \mathbb{R}^4$, the model of a three-layer neural network with input $x \in \mathbb{R}$, output $y \in \mathbb{R}$, and activation function tanh is defined as follows:

$$f(x, \theta) := w_{31} \tanh(w_1 x) + w_{32} \tanh(w_2 x).$$

Definition 7
The conditional probability $p(y|x, \theta)$ that the function approximation model Y follows is defined as the statistical model as follows:

$$p(y|x, \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{|y - f(x, \theta)|^2}{2\sigma^2} \right).$$

If the statistical model realizes the true distribution, the conditional probability $q(y|x)$ that the output Y follows is determined as a true distribution as follows:

$$q(y|x) = p(y|x, \theta_0) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{|y - f(x, \theta_0)|^2}{2\sigma^2} \right).$$

Definition 8
For field k, let $f \in k[x_1, x_2, \cdots, x_d]$ be a nonzero polynomial. Let the term of f that has the max monomial order be the leading term and denote it as $LT(f)$.

Definition 9
We fix the monomial order. Let the Gröbner basis be the finite subset $G = \{g_1, \cdots, g_t\}$ of ideal I that satisfies the following:

$$\langle LT(g_1), \cdots, LT(g_t) \rangle = \langle LT(I) \rangle.$$

Theorem 10 (Hilbert's basis theorem)
For an arbitrary ideal $I \subset k[x_1, x_2, \cdots, x_d]$, there exists a finitely generated set.

Definition 11
For ideal $I = \langle f_1, \cdots, f_s \rangle \subset k[x_1, x_2, \cdots, x_d]$, let the elimination ideal of the l-degree ideal of $k[x_{l+1}, \cdots, x_d]$ be the ideal that satisfies the following:

$$I_l := I \cap k[x_{l+1}, \cdots, x_d].$$

Theorem 12 (Elimination theorem)
Let $I \subset k[x_1, x_2, \cdots, x_d]$ be an ideal, and G be a Gröbner basis for the lexicographic order. For $0 \leq l \leq d$, the set G_l defined below is a Gröbner basis of an elimination ideal of l degree:

$$G_l := G \cap k[x_{l+1}, \cdots, x_d].$$

2 The set of parameters by which the true density function is realizable by the statistical mode: The case of $H = 2 \Rightarrow H_0 = 1$

Let the statistical model be a three-layer neural network with one input unit, $H = 1$ hidden units, and one output unit. Let the activation function be defined by the tanh. Additionally, let the true density function be a three layer neural network with one input unit, $H_0 = 0$ hidden units, and one output unit.

The $f(x, w)$ and true distribution are defined as follows:

$$f(x, w) = w_{31} \tanh(w_1 x) + w_{32} \tanh(w_2 x), \quad q(x, y) = \frac{q(x)}{\sqrt{2\pi}} \exp(-\frac{|y - e \tanh(fx)|^2}{2}).$$

Figure1 shows the case where the true distribution is realized with $H_0 = 1$.

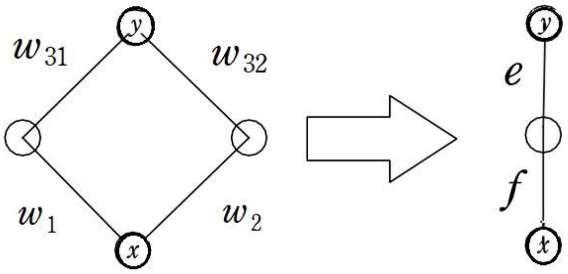


Figure 1: $H = 2 \Rightarrow H_0 = 1$

2.1 Set of parameters for which a true density function is realizable

We consider the analytic set of parameters by which the true density function is realizable by the statistical model.

$$W_0 := \{\theta \in \mathbb{R}^4 | p(x, y|\theta) = q(x, y)\} = \{\theta \in \mathbb{R}^4 | w_{31} \tanh(w_1 x) + w_{32} \tanh(w_2 x) = e \tanh(fx)\}.$$

Using Taylor expansion of tanh, we defined as follows:

$$g_k(w_{31}, w_1, w_{32}, w_2, e, f) := w_{31}w_1^{2k+1} + w_{32}w_2^{2k+1} - ef^{2k+1}.$$

The defining equation is represented as follows: (B_n denotes the Bernoulli numbers.)

$$\sum_{k=0}^{\infty} \frac{2^{2k+2}(2^{2k+2}-1)B_{2k+2}x^{2k+1}}{(2k+2)!} g_k(w_{31}, w_1, w_{32}, w_2, e, f).$$

$\{x^{2k+1}\}$is linearly independent, W_0 is a common zero point defined by infinite polynomials.

$$W_0 := \{\theta \in \mathbb{R}^4 | g_0 = g_1 = \cdots = 0\}.$$

If $I_k = \langle g_0, g_1, g_2, \cdots, g_k\rangle$ then, The nondecreasing sequence of ideals $I_0 \subset I_1 \subset I_2 \subset I_3 \subset \cdots$ stops and the following holds [5].

$$\begin{aligned} V(I_k) \quad &= \{\theta \in \mathbb{R}^4 | g_0 = g_1 = \cdots = g_k = 0\} \\ &= \{\theta \in \mathbb{R}^4 | w_{31}w_1 + w_{32}w_2 - ef = w_{31}w_1^3 + w_{32}w_2^3 - ef^3 = w_{31}w_1^5 + w_{32}w_2^5 - ef^5 = 0\}. \end{aligned}$$

2.2 Computation of algebraic sets

The coordinate transformation from the parameters $\theta = (w_1,\, w_2,\, w_{31},\, w_{32})$ to the new parameters $\xi = (a,\, b,\, v,\, w)$ is defined as follows [1].

$$a = w_2 - w_1,\ b = \frac{w_{31} - w_{32}}{w_{31} + w_{32}},\ v = \frac{w_{31}w_1 + w_{32}w_2}{w_{31} + w_{32}},\ w = w_{31} + w_{32}.$$

The inverse transformation is expressed as

$$w_1 = v + \frac{1}{2}a(b-1),\ w_2 = v + \frac{1}{2}a(b+1),\ w_{31} = \frac{1}{2}w(1+b),\ w_{32} = \frac{1}{2}w(1-b).$$

Using Mathematica, a simplified transformation from the coordinate system $\theta = (w_1,\, w_2,\, w_{31},\, w_{32})$ to $\xi = (a,\, b,\, v,\, w)$ yields the following output.

$$\begin{aligned} &f_1 = w_{31}w_1 + w_{32}w_2 - ef = vw - ef, \\ &f_2 = w_{31}w_1^3 + w_{32}w_2^3 - ef^3 = \tfrac{1}{4}w\left(a^3\left(b - b^3\right) - 3a^2\left(b^2 - 1\right)v + 4v^3\right) - ef^3, \\ &f_3 = w_{31}w_1^5 + w_{32}w_2^5 - ef^5 = \tfrac{1}{2}\left((b+1)w\left(\tfrac{1}{2}a(b-1) + v\right)^5 + (1-b)w\left(\tfrac{1}{2}a(b+1) + v\right)^5 - 2ef^5\right). \end{aligned}$$

We then input the following to calculate the Gröbner basis of the ideal eliminated v, w:

```
GroebnerBasis[{f1, f2, f3}, {a,b,v,w,e,f}, {v,w}, MonomialOrder -> Lexicographic].
```

This can be factored as follows:

$$a^2(-1+b)(1+b)e(a-2f)^2(a-f)f(a+f)(a+2f)^2.$$

Thus $a = 0$ or $b = \pm 1$ or $a = \pm f$ or $a = \pm 2f$.

2.2.1 Parameter representation of algebraic set (1)

(1) When $a = 0, b = \pm 1$, using Mathematica, a simplified transformation from the coordinate system $\theta = (w_1,\, w_2,\, w_{31},\, w_{32})$ to $\xi = (a,\, b,\, v,\, w)$ yields the following output.

$$f_1 = vw - ef,\; f_2 = v^3 w - ef^3,\; f_3 = v^5 w - ef^5.$$

We then input the following to calculate the Gröbner basis of the ideal eliminated w:

```
GroebnerBasis[{f2, f3}, {v,w,e,f}, {w}, MonomialOrder -> Lexicographic].
```

This can be factored as follows: $-ef^3(f-v)(f+v)$.

Then the following holds: $f - v = 0$ or $f + v = 0$. Thus $v = \pm f$.
When $v = \pm f$, the following holds: $f_2 = f^3(w \mp e)$, $f_3 = f^5(w \mp e)$.

Thus $w = \pm e$. This solution holds $f_1 = 0$.

Therefore, the set of parameters that realize the true distribution is expressed as follows:

$$(0, b, \pm f, \pm e),\; (a, 1, \pm f, \pm e),\; (a, -1, \pm f, \pm e).$$

2.2.2 Parameter representation of algebraic set (2)

(2) When $a = \pm f$, using Mathematica, a simplified transformation from the coordinate system $\theta = (w_1,\, w_2,\, w_{31},\, w_{32})$ to $\xi = (a,\, b,\, v,\, w)$ yields the following output.

$$\begin{aligned}
f_1 &= vw - ef,\\
f_2 &= \tfrac{1}{4} w \left(\mp b^3 f^3 - 3b^2 f^2 v \pm bf^3 + 3f^2 v + 4v^3\right) - ef^3,\\
f_3 &= \tfrac{1}{2}\left((\pm b \pm 1) w \left(\tfrac{1}{2}(b-1)f \pm v\right)^5 \pm (1 \mp b) w \left(\tfrac{1}{2}(b+1)f \pm v\right)^5 - 2ef^5\right).
\end{aligned}$$

We then input the following to calculate the Gröbner basis of the ideal eliminated w:

```
GroebnerBasis[{f2, f3}, {b,v,w,e,f}, {w}, MonomialOrder -> Lexicographic].
```

This can be factored as follows:

$$ef^3(bf - f \pm 2v)(bf + f \pm 2v)\left(2b^3 f^3 \pm 7b^2 f^2 v - 2bf^3 + 4bfv^2 \mp 7f^2 v \mp 4v^3\right).$$

Then the following holds: $bf - f \pm 2v = 0$ or $bf + f \pm 2v = 0$. Thus $v = \pm\frac{(1-b)f}{2}$ or $v = \mp\frac{(1+b)f}{2}$.

(i) When $v = \pm\frac{(1-b)f}{2}$, the following holds: $f_2 = \frac{1}{2} f^3(\mp bw - 2e \pm w)$, $f_3 = \frac{1}{2} f^5(\mp bw - 2e \pm w)$.
Then the following holds: $\mp bw - 2e \pm w = 0$. Thus $w = \pm\frac{2e}{1-b}$. This solution holds $f_1 = 0$.

(ii) When $v = \mp\frac{(1+b)f}{2}$, the following holds: $f_2 = -\frac{1}{2} f^3(\pm bw + 2e \pm w)$, $f_3 = -\frac{1}{2} f^5(\pm bw + 2e \pm w)$.
Then the following holds: $\pm bw + 2e \pm w = 0$. Thus $w = \mp\frac{2e}{1+b}$. This solution holds $f_1 = 0$.

Therefore, the set of parameters as follows:

$$(\pm f, b, \pm\tfrac{(1-b)f}{2}, \pm\tfrac{2e}{1-b}),\; (\pm f, b, \mp\tfrac{(1+b)f}{2}, \mp\tfrac{2e}{1+b}).$$

2.2.3 Parameter representation of algebraic set (3)

(3) When $a = \pm 2f$, using Mathematica, a simplified transformation from the coordinate system $\theta = (w_1,\, w_2,\, w_{31},\, w_{32})$ to $\xi = (a,\, b,\, v,\, w)$ yields the following output.

$$\begin{aligned} f_1 &= vw - ef, \\ f_2 &= w\left(\mp 2b^3f^3 - 3b^2f^2v \pm 2bf^3 + 3f^2v + v^3\right) - ef^3, \\ f_3 &= \tfrac{1}{2}\left(\pm(b+1)w((b-1)f \pm v)^5 \mp (b-1)w((b+1)f \pm v)^5 - 2ef^5\right). \end{aligned}$$

We then input the following to calculate the Gröbner basis of the ideal eliminated w:

```
GroebnerBasis[{f2, f3}, {b,v,w,e,f}, {w}, MonomialOrder -> Lexicographic].
```

This can be factored as follows:

$$ef^3(bf \pm v)\left(4b^4f^4 \pm 11b^3f^3v - 2b^2f^4 + 9b^2f^2v^2 - 11bf^3v + bfv^3 - 2f^4 - 9f^2v^2 - v^4\right).$$

Then the following holds: $bf \pm v = 0$. Thus $v = \mp bf$.
When $v = \mp bf$, the following holds: $f_2 = \mp f^3(bw \pm e)$, $f_3 = \mp f^5(bw \pm e)$.

Then the following holds: $bw \pm e = 0$. Thus $w = \mp \frac{e}{b}$. This solution holds $f_1 = 0$.

Therefore, the set of parameters that realize the true distribution is expressed as follows:

$$(\pm 2f, b, \mp bf, \mp \tfrac{e}{b}).$$

2.2.4 Representation of algebraic sets from (1) to (3)

Thus if the true distribution is realized with $H_0 = 1$ the set of parameters that realize the true distribution is expressed as follows (congruent order):

$$\begin{aligned} &(1)(0, b, \pm f, \pm e),\ (a, 1, \pm f, \pm e),\ (a, -1, \pm f, \pm e), \\ &(2)(\pm f, b, \pm\tfrac{(1-b)f}{2}, \pm\tfrac{2e}{1-b}),\ (\pm f, b, \mp\tfrac{(1+b)f}{2}, \mp\tfrac{2e}{1+b}),\ (3)(\pm 2f, b, \mp bf, \mp\tfrac{e}{b}). \end{aligned}$$

2.3 Singular points of an algebraic set

Next, the fundamental transformation of the Jacobi matrix is performed and the rank of the Jacobi matrix is calculated for the singular points of the algebraic set. The Jacobi matrix A is defined as follows:

$$A = \begin{pmatrix} \frac{\partial f_1}{\partial a} & \frac{\partial f_1}{\partial b} & \frac{\partial f_1}{\partial v} & \frac{\partial f_1}{\partial w} \\ \frac{\partial f_2}{\partial a} & \frac{\partial f_2}{\partial b} & \frac{\partial f_2}{\partial v} & \frac{\partial f_2}{\partial w} \\ \frac{\partial f_3}{\partial a} & \frac{\partial f_3}{\partial b} & \frac{\partial f_3}{\partial v} & \frac{\partial f_3}{\partial w} \end{pmatrix}.$$

Here, define $f_1,\, f_2,\, f_3$ as follows:

$$\begin{aligned} &f_1 = vw - ef,\ f_2 = \tfrac{1}{4}w\left(a^3\left(b - b^3\right) - 3a^2\left(b^2 - 1\right)v + 4v^3\right) - ef^3, \\ &f_3 = \tfrac{1}{2}\left((b+1)w\left(\tfrac{1}{2}a(b-1) + v\right)^5 + (1-b)w\left(\tfrac{1}{2}a(b+1) + v\right)^5 - 2ef^5\right). \end{aligned}$$

Here, the following holds:

$$\begin{aligned}
&\tfrac{\partial f_1}{\partial a}=0,\ \tfrac{\partial f_1}{\partial b}=0,\ \tfrac{\partial f_1}{\partial v}=w,\ \tfrac{\partial f_1}{\partial w}=v,\\
&\tfrac{\partial f_2}{\partial a}=\tfrac{1}{4}(-3)a(b-1)(b+1)w(ab+2v),\ \tfrac{\partial f_2}{\partial b}=\tfrac{1}{4}a^2w\left(-3ab^2+a-6bv\right),\\
&\tfrac{\partial f_2}{\partial v}=-\tfrac{3}{4}w\left(a^2b^2-a^2-4v^2\right),\ \tfrac{\partial f_2}{\partial w}=\tfrac{1}{4}\left(-a^3b^3+a^3b-3a^2b^2v+3a^2v+4v^3\right),\\
&\tfrac{\partial f_3}{\partial a}=\tfrac{1}{8}(-5)a(b-1)(b+1)w(ab+2v)\left(a^2b^2+a^2+4abv+4v^2\right),\\
&\tfrac{\partial f_3}{\partial b}=\tfrac{1}{8}a^2w\left(a^3\left(1-5b^4\right)+10a^2b\left(1-3b^2\right)v+20a\left(1-3b^2\right)v^2-40bv^3\right),\\
&\tfrac{\partial f_3}{\partial v}=\tfrac{5}{2}w\left((b+1)\left(\tfrac{1}{2}a(b-1)+v\right)^4+(1-b)\left(\tfrac{1}{2}a(b+1)+v\right)^4\right),\\
&\tfrac{\partial f_3}{\partial w}=\tfrac{1}{2}\left((b+1)\left(\tfrac{1}{2}a(b-1)+v\right)^5+(1-b)\left(\tfrac{1}{2}a(b+1)+v\right)^5\right).
\end{aligned}$$

The singular point sets that realize the true distribution are analyzed in (1)-(4) below:

$$(1)(0,b,\pm f,\pm e),\ (2)(\pm 2f,b,\mp bf,\mp\tfrac{e}{b}),$$

$$(3)\left(\pm f,b,\pm\tfrac{(1-b)f}{2},\pm\tfrac{2e}{1-b}\right),\ \left(\pm f,b,\mp\tfrac{(1+b)f}{2},\mp\tfrac{2e}{1+b}\right),\ (4)(a,-1,\pm f,\pm e),\ (a,1,\pm f,\pm e).$$

2.3.1 Representation of singular sets (1)

(1) When $(0,b,\pm f,\pm e),\ A=\begin{pmatrix}0&0&\pm e&\pm f\\0&0&\pm 3ef^2&\pm f^3\\0&0&\pm 5ef^4&\pm f^5\end{pmatrix}\longrightarrow\begin{pmatrix}0&0&0&1\\0&0&1&0\\0&0&0&0\end{pmatrix}$. Thus, rank $A=2$ holds.

2.3.2 Representation of singular sets (2)

(2) When $(\pm 2f,b,\mp bf,\mp\frac{e}{b}),\ A=\begin{pmatrix}0&0&\mp\frac{e}{b}&\mp bf\\0&-\frac{2ef^3}{b}&\mp\frac{3ef^2}{b}&\mp bf^3\\0&-\frac{4ef^5}{b}&\mp\frac{5ef^4}{b}&\mp bf^5\end{pmatrix}\longrightarrow\begin{pmatrix}0&0&0&1\\0&1&0&0\\0&0&0&0\end{pmatrix}$. Thus, rank $A=2$ holds.

2.3.3 Representation of singular sets (3)

(3) When $(\pm f,b,\pm\frac{(1-b)f}{2},\pm\frac{2e}{1-b}),\ A=\begin{pmatrix}0&0&\mp\frac{2e}{b-1}&\mp\frac{(b-1)f}{2}\\\pm\frac{3(b+1)ef^2}{2}&\frac{(3b-1)ef^3}{2(b-1)}&\pm 3ef^2&\mp\frac{(b-1)f^3}{2}\\\pm\frac{5(b+1)ef^4}{2}&\frac{(5b-3)ef^5}{2(b-1)}&\pm 5ef^4&\mp\frac{(b-1)f^5}{2}\end{pmatrix}$.

(i) When $b\neq -1,\ A\longrightarrow\begin{pmatrix}0&0&1&0\\0&1&0&0\\1&0&0&0\end{pmatrix}$. Thus, rank $A=3$ holds.

(ii) When $b=-1,\ A=\begin{pmatrix}0&0&\pm e&\pm f\\0&ef^3&\pm 3ef^2&\pm f^3\\0&2ef^5&\pm 5ef^4&\pm f^5\end{pmatrix}\longrightarrow\begin{pmatrix}0&0&1&0\\0&1&0&0\\0&0&0&0\end{pmatrix}$. Thus, rank $A=2$ holds.

Therefore, rank $A=\begin{cases}3 & (b\neq -1)\\2 & (b=-1)\end{cases}$ holds.

When $(\pm f, b, \mp\frac{(1+b)f}{2}, \mp\frac{2e}{1+b})$, $A = \begin{pmatrix} 0 & 0 & \mp\frac{2e}{b+1} & \mp\frac{(b+1)f}{2} \\ \mp\frac{3(b-1)ef^2}{2} & -\frac{(3b+1)ef^3}{2(b+1)} & \mp 3ef^2 & \mp\frac{(b+1)f^3}{2} \\ \mp\frac{5(b-1)ef^4}{2} & -\frac{(5b+3)ef^5}{2(b+1)} & \mp 5ef^4 & \mp\frac{(b+1)f^5}{2} \end{pmatrix}$.

(i) When $b \neq 1$, $A \longrightarrow \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$. Thus, rank $A = 3$ holds.

(ii) When $b = 1$, $A = \begin{pmatrix} 0 & 0 & \mp e & \mp f \\ 0 & -ef^3 & \mp 3ef^2 & \mp f^3 \\ 0 & -2ef^5 & \mp 5ef^4 & \mp f^5 \end{pmatrix} \longrightarrow \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$. Thus, rank $A = 2$ holds.

Therefore, rank $A = \begin{cases} 3 & (b \neq 1) \\ 2 & (b = 1) \end{cases}$ holds.

2.3.4 Representation of singular sets (4)

(4) When $(a, 1, \pm f, \pm e)$, $A = \begin{pmatrix} 0 & 0 & \pm e & \pm f \\ 0 & \mp\frac{1}{2}a^2e(a \pm 3f) & \pm 3ef^2 & \pm f^3 \\ 0 & \mp\frac{1}{2}a^2e\left(a^3 \pm 5a^2f + 10af^2 \pm 10f^3\right) & \pm 5ef^4 & \pm f^5 \end{pmatrix}$.

(i) When $a \neq 0$, $A \longrightarrow \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & (a \pm f)(a \pm 2f)^2 & 0 & 0 \end{pmatrix}$. Thus, rank $A = \begin{cases} 3 & (a \neq \mp f,\ \mp 2f) \\ 2 & (a = \mp f,\ \mp 2f) \end{cases}$ holds.

(ii) When $a = 0$, rank $A = 2$ holds as in (1).

Therefore, rank $A = \begin{cases} 3 & (a \neq \mp f,\ \mp 2f, 0) \\ 2 & (a = \mp f,\ \mp 2f, 0) \end{cases}$ holds.

When $(a, -1, \pm f, \pm e)$, $A = \begin{pmatrix} 0 & 0 & \pm e & \pm f \\ 0 & \mp\frac{1}{2}a^2e(a \mp 3f) & \pm 3ef^2 & \pm f^3 \\ 0 & \mp\frac{1}{2}a^2e\left(a^3 \mp 5a^2f + 10af^2 \mp 10f^3\right) & \pm 5ef^4 & \pm f^5 \end{pmatrix}$.

(i) When $a \neq 0$, $A \longrightarrow \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & (a \mp f)(a \mp 2f)^2 & 0 & 0 \end{pmatrix}$. Thus, rank $A = \begin{cases} 3 & (a \neq \pm f,\ \pm 2f) \\ 2 & (a = \pm f,\ \pm 2f) \end{cases}$ holds.

(ii) When $a = 0$, rank $A = 2$ holds as in (1).

Therefore, rank $A = \begin{cases} 3 & (a \neq \mp f,\ \mp 2f, 0) \\ 2 & (a = \mp f,\ \mp 2f, 0) \end{cases}$ holds.

2.3.5 Representation of singular sets from (1) to (4)

Thus if the true distribution is realized with $H_0 = 1$ the set of parameters that realize the true distribution is expressed as follows(congruent order):

$$(1)(0, b, \pm f, \pm e),\ (2)(\pm 2f, b, \mp bf, \mp\frac{e}{b}),\ (3)(\pm f, -1, \pm f, \pm e),\ (\pm f, 1, \mp f, \mp e),$$
$$(4)(\mp f, 1, \pm f, \pm e),\ (\mp 2f, 1, \pm f, \pm e),\ (0, 1, \pm f, \pm e),\ (\mp f, -1, \pm f, \pm e),\ (\mp 2f, -1, \pm f, \pm e),\ (0, -1, \pm f, \pm e).$$

The region represented by $a = 0,\ b = \pm 1$ is discussed as follows.

3 Singular region

3.1 Overlap singularity, Elimination singularity

Definition 13

An overlap singularity is defined as the region in the parameter space where θ and ξ satisfy

$$R_0 := \{\theta \in \mathbb{R}^4 | w_1 = w_2\} = \{\xi \in \mathbb{R}^4 | a = 0\}.$$

An elimination singularity is defined as the region in the parameter space where θ and ξ satisfy

$$R_1 := \{\theta \in \mathbb{R}^4 | w_{31} = 0\} \cup \{\theta \in \mathbb{R}^4 | w_{32} = 0\} = \{\xi \in \mathbb{R}^4 | b = \pm 1\}.$$

Consider the learning dynamics in the neighborhood of $a = 0, b = \pm 1$.

3.2 Learning function approximation models

Updating the current estimate to a new estimate for the observed data to minimize loss is called learning in information science. The function approximation model is trained to find ξ that minimizes the square error $l(x, y, \xi) = \frac{1}{2}|y - f(x, \xi)|^2$.

(1) When $a \approx 0,$ since l_v, l_w is of order $O(1)$, the two parameters $(v,\ w)$ quickly reach equilibrium when learning begins [3].

(2) When $a \approx 0,$ since l_a is of order $O(a)$, l_b is of order $O(a^2)$, $(a,\ b)$ changes slowly in this state [3].

We fix $(v,\ w)$ as the optimal solution (v^*, w^*) and stop changing the parameters $(v,\ w)$, and consider the dynamics of the parameters $(a,\ b)$ to the optimal solution.

3.3 Construction of neural networks

Let a, b be variables and c, d, v, w be constants. Using Mathematica, $F1$, $F2$, $elem1$, $elem3$, $elem4$, $elem0$, $elem2$ define as follows:

```
F1[a_ ] := NetInsertSharedArrays[ NetChain [LinearLayer[1, "Weights" -> a, "Biases" -> None]], "Linear1"],
F2[b_] := NetInsertSharedArrays[ NetChain [LinearLayer[1, "Weights" -> b, "Biases" -> None]], "Linear2"],
elem1[v_, d_ ] := ElementwiseLayer[#*(v) + d &],
elem3[w_ ] := ElementwiseLayer[#*(w) &], elem4[c_ ] := ElementwiseLayer[# + (c) &],
elem0 := ElementwiseLayer[#*(1/2) &], elem2 := ElementwiseLayer[#*(-1) &].
```

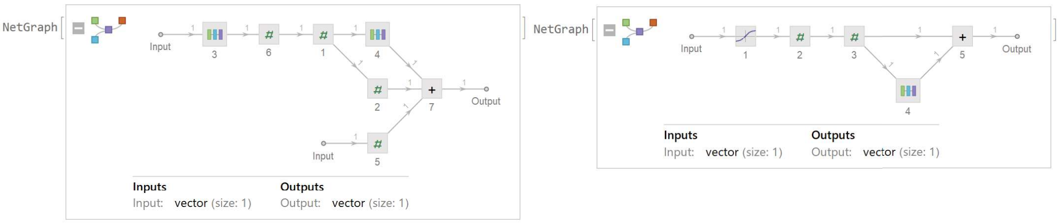
First, input condition $w_1 x = \left(v + \frac{1}{2}(b-1)a\right) x$ as follows:

```
net11[a_, b_, c_, d_, v_ ] := NetGraph[elem0, elem2, F1[a], F2[b], elem1[v, d], elem4[c],
TotalLayer[], NetPort["Input"]-> 3, 3 -> 6 -> 1, 1 -> 4,1 -> 2, {4, 2, 5} -> 7].
```

Input next condition $w_{31} x = \frac{1}{2} w(b+1)\tanh(x)$ as follows:

```
net12[a_, b_, c_, d_, w_ ] := NetGraph[Tanh, elem0, elem3[w], F2[b],TotalLayer[],
NetPort["Input"] -> 1, 1 -> 2 -> 3 -> 4, {3, 4} -> 5].
```

*Net*11 is shown on the left side of Figure 2 and *net*12 is shown on the right side of Figure 2.
Input next condition $w_{31}\tanh(w_1 x) = \frac{1}{2} w(b+1)\tanh[\left(v + \frac{1}{2}(b-1)a\right) x]$ as follows:



Inputs
Input: vector (size: 1)

Outputs
Output: vector (size: 1)

Inputs
Input: vector (size: 1)

Outputs
Output: vector (size: 1)

Figure 2: *net*11, *net*12

```
net1[a_, b_, c_, d_, v_, w_ ] := NetGraph[net11[a, b, c, d, v], net12[a, b, c, d, w],
NetPort["Input"] -> 1, 1 -> 2].
```

Similarly, *net*21, *net*22, *net*2 are defined to express the condition.

$$w_2 x = \left(v + \tfrac{1}{2}(b+1)a\right) x,\ w_{32} x = \tfrac{1}{2} w(-b+1)\tanh(x).$$

Finally, input next condition $w_{31}\tanh(w_1 x) + w_{32}\tanh(w_2 x)$ as follows:

```
parameterNet[a_, b_, c_, d_, v_, w_ ] := NetGraph[net1[a, b, c, d, v, w],
net2[a, b, c, d, v, w], TotalLayer[], NetPort["Input"] -> 1, NetPort["Input"] -> 2,
{1, 2} -> 3 -> NetPort["Output1"], "Input" -> enc].
```

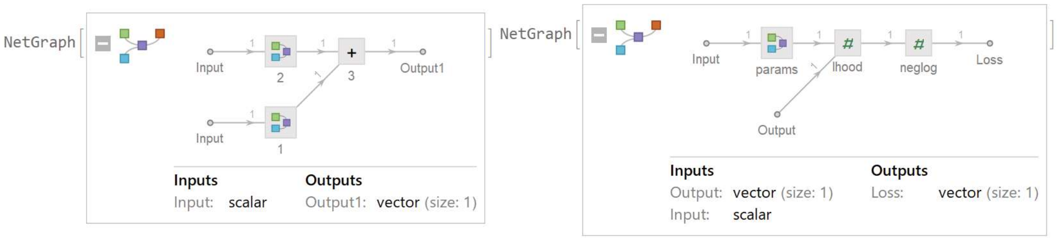
Using Mathematica, a gaussian function is defined as follows:

```
gaussianLikelihood[y\_, \mu \_] := PDF[NormalDistribution[\mu, 1], y].
```

The estimation method with the squared error as the loss function corresponds to the maximum likelihood estimation method for $p(y|x,\xi)$. A NetGraph is constructed with the loss function as the log likelihood ratio function to define *trainingNet*. We input the following:

```
trainingNet[a_, b_, v_, w_] := NetGraph[<|"params" -> parameterNet[a, b, v, w],
"lhood" ->ThreadingLayer[gaussianLikelihood],"neglog" -> ElementwiseLayer[-Log[#]&]|>,
NetPort["Output"], NetPort["params", "Output1"] ->"lhood", "lhood" -> "neglog" -> NetPort["Loss"]].
```

ParameterNet is shown on the left side of Figure 3 and *trainingNet* is shown on the right side of Figure 3.


NetGraph

Inputs
Input: scalar

Outputs
Output1: vector (size: 1)

NetGraph

Inputs
Output: vector (size: 1)
Input: scalar

Outputs
Loss: vector (size: 1)

Figure 3: $parameterNet$, $trainingNet$

To define the loss function $L_n(\xi) := \frac{1}{2n}\sum_{i=1}^{n} |y_i - f(X_i, \xi)|^2$, we input the following:

```
G[a_, b_ ] := Mean[trainingNet[a, b, c0, d0, v0, w0] [<|"Input" -> dataX, "Output" -> enc[dataY]|>]].
```

4 Application to mathematics learning

As an example of application, overgeneralization (the phenomenon of overgeneralizing specific rules or semantic features) may occur in the process of learners gaining an understanding of the two concepts in relation to each other [7], [8], [9]. The phenomenon of overgeneralization is represented in the figure based on the learner's test scores and the method using neural networks in information science. Overview of applications to mathematics learning is shown on Figure 4.

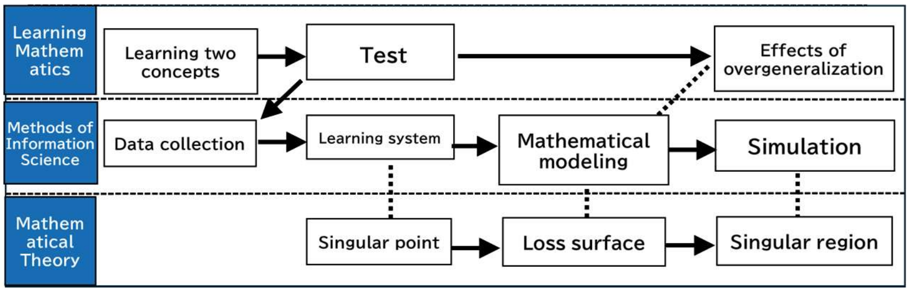


Figure 4: Overview of applications to mathematics learning

The dynamics of learning near singularity is classified into following five patterns by changing an initial value [4].

(1) Fast convergence : The learning process converges to the global minimum fast.

(2) Overlap singularity : The learning process is significantly affected by overlap singularity.

(3) Cross elimination singurarity : The learning process crosses the elimination and reaches the global optimum after training.

(4) Near elimination singularity : The learning process is significantly affected by elimination singularity.

(5) Output weight 0 : After training, output weight becomes nearly equal to 0.

An example of a simulation method that visualizes the learner's state of understanding on a loss surface. The dynamics of learning near the singular region can be used to determine whether varying the degree of overgeneralization will advance understanding, and whether changing the proportion of the learner population will reduce overgeneralization. An example of a simulation is shown on Figure 5.

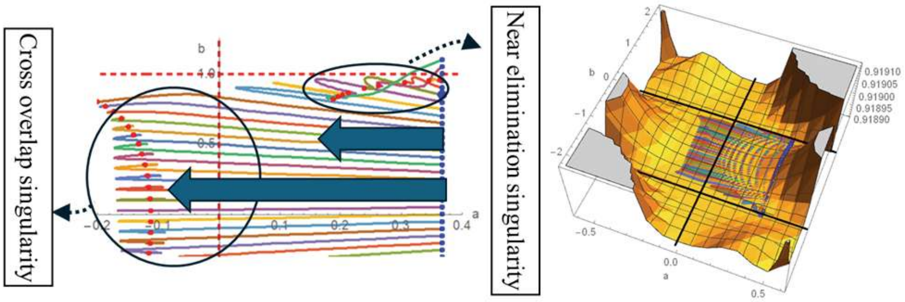


Figure 5: Example of a simulation

5 Future issues

- Generalize the number of hidden units in the statistical model to obtain parameter representations in the coordinate system ξ of the algebraic sets and singular sets that realize the true distribution.
- Analyze the structure of singularities by using Mathematica and find conditions for stable learning from learning equations, and clarify aspects of understanding in mathematics education.

References

[1] S. Amari, “Information geometry and Its applications,” Springer, 2016.

[2] K. Fukumizu and S. Amari, “Local minima and plateaus in hierarchical structures of multilayer perceptrons,” Neural Netw, vol. 13, pp. 317–327, 2000.

[3] H. Wei, J. Zhang, F. Cousseau, T. Ozeki, and S. Amari, “Dynamics of learning near singularities in layered networks,” Neural Computation, vol. 20, No. 34, pp. 813-843, 2008.

[4] W. Guo, H. Wei , Y. Ong, J. R. Hervas, J. Zhao, H. Wang, K. Zhang, “Numerical Analysis near Singularities in RBF Networks,” The Journal of Machine Learning Research, 19(1), pp. 1 − 39, 2018.

[5] S. Watanabe, “Algebraic geometry and statistical learning theory,” Cambridge University Press, 2009.

[6] S. Watanabe, “Mathematical theory of bayesian statistics,” CRC Press, 2018.

[7] T. Washino and S. Ohashi, “Learning guidance based on the overlap singularity phenomenon,” Scientiae Mathematicae Japonicae, 31572, pp. 1-20, 2023.

[8] T. Washino and T. Takahashi, “Learning guidance based on the elimination singularity phenomenon,” Proceedings of 28th Asian Technology Conference in Mathematics, pp. 352-361, 2023.

[9] T. Washino and T. Takahashi, “Learning guidance based on analysis of symbolic comprehension,” Proceedings of 29th Asian Technology Conference in Mathematics, pp. 200-209, 2024.