On the analysis of singularity structure in learning

*Tomohiro Washino**, *Tadashi Takahashi*
∗Corresponding Author: d1523001@s.konan-u.ac.jp,
takahasi@konan-u.ac.jp
Department of Intelligence and Informatics
Konan University
8-9-1 Okamoto, Higashinada, Kobe
Japan

Abstract

The existence of singularities often affects the learning dynamics in neural network and caused plateau phenomena. Using Mathematica, we observed near singular regions by examining the evolution of the parameter and the dynamics of learning on the training loss surface. Our result is to investigate that the type of dynamics of learning changes when the overlap and the elimination singularity is approached from a distance by changing the initial values of the statistical model, and to clarify the plateau phenomenon observed near singular regions.

1 Introduction

In a hierarchical structure model which is a neural network, a set of true parameters consists of not a union of several manifolds. Watanabe[1], [2] investigated that the analytic set of parameters contains singularities by using algebraic geometry and Bayesian statistics.

Let the statistical model be a three-layer neural network, plateau phenomena were observed in singular regions where two hidden neurons can be rewritten with only one hidden neuron and pose a serious problem in neural networks[3], [4], [5]. Amari[5] showed that a subset of critical points corresponding to the global minimum of a smaller network can be local minima or saddles of the larger network. Amari[3] discussed the learning dynamics near the overlap singularity and the elimination singularity close to them. Also, Amari[4] introduced coordinate transformation of parameters of the statistical model and fixed variables moving quickly and searched trajectories of learning of variables moving slowly. Moreover, he[5] calculated stability and dynamics of learning near singular regions. Guo[6] classified dynamics of learning near the overlap singularity and the elimination singularity into five patterns.

Currently, the dynamics of learning near the elimination singularities far away from the overlap singularities still remains unknown. By continuously changing the initial value of learning, we investigate that the type of dynamics according to Guo's classification of overlap and elimination singularity when approaching from a distance changes.

2 Definition

Definition 1 (Input, noise, training data, test data) *We assume that an $\mathbb{R}^1$-valued random variable X that follows a probability density function p(x) is input and that an $\mathbb{R}^1$-valued random variable Z that follows a normal distribution, the average and standard derivation of which are $(0,\ \sigma)$, is noise. For $\theta_{\mathbf{0}} = (w^*_{11},\ w^*_{12},\ w^*_{21},\ w^*_{22},\ w^*_{31},\ w^*_{32}) \in \mathbb{R}^6$, an $\mathbb{R}^1$-valued random variable Y is determined by the training data or test data as follows[1], [2]:*

$$Y := f(x, \theta_0) + Z = w^*_{31} \tanh(w^*_{11} x + w^*_{21}) + w^*_{32} \tanh(w^*_{12} x + w^*_{22}) + Z.$$

Definition 2 (Function approximation model) *For parameters $\theta = (\mathbf{w_1},\ \mathbf{w_2},\ w_{31},\ w_{32}) \in \mathbb{R}^6$, and $\mathbb{R}^1$-valued function $f(x, \theta)$, an $\mathbb{R}^1$-valued random variable Y is determined as a function approximation model as follows[1], [2]:*

$$\begin{aligned} Y &:= f(x,\theta) + Z = w_{31}\phi(\mathbf{x}, \mathbf{w_1}) + w_{32}\phi(\mathbf{x}, \mathbf{w_2}) + Z = w_{31}\tanh(\mathbf{w_1}^{\mathrm{T}}\mathbf{x}) + w_{32}\tanh(\mathbf{w_2}^{\mathrm{T}}\mathbf{x}) + Z \\ &= w_{31}\tanh(w_{11}x + w_{21}) + w_{32}\tanh(w_{12}x + w_{22}) + Z. \end{aligned}$$

where $\mathbf{w_1} = (w_{11},\ w_{12})$, $\mathbf{w_2} = (w_{21},\ w_{22})$, $\mathbf{x} = (x,\ 1)$.

Definition 3 (Statistical model, true density function) *A conditional probability density that follows function approximation model Y and is referred to as a statistical model is defined as follows[1], [2]:*

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

A conditional probability density that follows output Y and is referred to as a true density function is defined as follows[1], [2]:

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

Definition 4 (Overlap singularity,elimination singularity) *An overlap singularity is defined as the special region in the parameter space in which w_i satisfies[3]*

$$R_0 := \{\theta \in \mathbb{R}^6 | \mathbf{w_1} = \mathbf{w_2}\}.$$

The elimination singularity is defined as the special region in the parameter space in which w_i satisfies[3]

$$R_1 := \{\theta \in \mathbb{R}^6 | w_{31} = 0\} \cup \{\theta \in \mathbb{R}^6 | w_{32} = 0\}.$$

We recall the following coordinate transformation from the parameter $\theta = (\mathbf{w_1},\ \mathbf{w_2},\ w_{31},\ w_{32})$ to the parameter $\xi = (\mathbf{a},\ b,\ \mathbf{v},\ w)$[3]:

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

Using coordinate ξ, the coordinate θ is as follows :

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

For $y = f(x, \theta_0) + Z,$we define the loss function

$$l(y, \mathbf{x}, \theta) := \frac{1}{2}(y - f(\mathbf{x}, \theta))^2.$$

Then, for a learning rateη, parameter θ, which is modified by the stochastic gradient descent algorithm, is as follows:

$$\theta(t+1) - \theta(t) := -\eta \frac{\partial l(y_t, \mathbf{x_t}, \theta_t)}{\partial \theta}.$$

Definition 5 (Learning equation of coordinate θ) *For coordinate* $\theta = (\mathbf{w_1},\ \mathbf{w_2},\ w_{31},\ w_{32})$, *the learning equation is defined as follows:*

$$\dot{\theta}(t) := -\eta \left\langle \frac{\partial l(y, \mathbf{x}, \theta)}{\partial \theta} \right\rangle = \int -\eta \frac{\partial l(y, \mathbf{x}, \theta)}{\partial \theta} q(y|\mathbf{x}) dy d\mathbf{x}.$$

For loss $e(y, \mathbf{x}, \xi) := y - f(\mathbf{x}, \xi)$, negative gradients of the loss function $l(\xi)$ hold as follows[1]:

$$\begin{aligned}
l_{\mathbf{v}}(\xi) &= w \left\langle e(y, \mathbf{x}, \xi) \frac{\partial \phi(\mathbf{x}, \mathbf{v})}{\partial \mathbf{v}} \right\rangle + \frac{1}{8} w (1 - z^2) Q(\mathbf{v}, \mathbf{a}) + O(\mathbf{a}^3), \\
l_w(\xi) &= \langle e(y, \mathbf{x}, \xi) \phi(\mathbf{x}, \mathbf{v}) \rangle + \frac{1}{8}(1 - z^2) \left\langle e(y, \mathbf{x}, \xi) \mathbf{a}^{\mathrm{T}} \frac{\partial^2 \phi(\mathbf{x}, \mathbf{v})}{\partial \mathbf{v} \partial \mathbf{v}^{\mathrm{T}}} \mathbf{a} \right\rangle + O(\mathbf{a}^3), \\
l_{\mathbf{a}}(\xi) &= \frac{1}{4} w (1 - z^2) \left\langle e(y, \mathbf{x}, \xi) \mathbf{a} \frac{\partial^2 \phi(\mathbf{x}, \mathbf{v})}{\partial \mathbf{v} \partial \mathbf{v}^{\mathrm{T}}} \right\rangle + \frac{1}{24} w z (1 - z^2) \left\langle e(y, \mathbf{x}, \xi) \frac{\partial D(x, \mathbf{v}, \mathbf{a})}{\partial \mathbf{a}} \right\rangle + O(\mathbf{a}^3), \\
l_b(\xi) &= -\frac{1}{4} w z \left\langle e(y, \mathbf{x}, \xi) \mathbf{a}^{\mathrm{T}} \frac{\partial^2 \phi(\mathbf{x}, \mathbf{v})}{\partial \mathbf{v} \partial \mathbf{v}^{\mathrm{T}}} \mathbf{a} \right\rangle + O(\mathbf{a}^3).
\end{aligned}$$

where $Q(v, \mathbf{a}) := \left\langle e(y, \mathbf{x}, \xi) \frac{\partial}{\partial \mathbf{v}} (\mathbf{a}^{\mathrm{T}} \frac{\partial^2 \phi(\mathbf{x}, \mathbf{v})}{\partial \mathbf{v} \partial \mathbf{v}^{\mathrm{T}}} \mathbf{a}) \right\rangle,\ D(\mathbf{x}, \mathbf{v}, \mathbf{a}) := \sum_{i,\,j,\,k} \frac{\partial^3 \phi(\mathbf{x}, \mathbf{v})}{\partial v_i \partial v_j \partial v_k} a_i a_j a_k.$

Note that $l_{\mathbf{v}},\ l_w$is of order $O(1)$. By taking into account the fact that $\mathbf{a} \approx \mathbf{0},$ we see that the time evolution of $(\mathbf{v},\ w)$ is fast and converges to the partial equilibrium states that satisfies $l_{\mathbf{v}}(\xi) = l_w(\xi) = 0$ quickly.

On the other hand, note that $l_{\mathbf{a}}$ and $l_{\mathbf{b}}$ is of order $O(\mathbf{a})$ and $O(\mathbf{a}^2)$. By taking into account the fact that $\mathbf{a} \approx \mathbf{0},$ we see that the time evolution of $(\mathbf{a},\ b)$ is slow[4].

Definition 6 (Learning equation of coordinate ξ) *For the coordinate* $\xi = (\mathbf{a},\ b,\ \mathbf{v},\ w)$, *the learning equation is defined as follows:*

$$\dot{\xi} := -\eta \frac{\partial \xi}{\partial \theta^{\mathrm{T}}} \left(\frac{\partial \xi}{\partial \theta^{\mathrm{T}}} \right)^{\mathrm{T}} \left\langle \frac{\partial l(y, \mathbf{x}, \xi)}{\partial \xi} \right\rangle.$$

Then, the learning equations hold as follows[3]:

$$\begin{aligned}
&\dot{\mathbf{v}} = \frac{b^2+1}{2} l_{\mathbf{v}} + \frac{b^2+1}{2w^2} \mathbf{a}\mathbf{a}^{\mathrm{T}} l_{\mathbf{v}} + \frac{b}{w} \mathbf{a} l_w - b l_{\mathbf{a}} - \frac{b^2+1}{w^2} \mathbf{a} l_b, \quad \dot{w} = \frac{b}{w} \mathbf{a}^{\mathrm{T}} l_{\mathbf{v}} + 2 l_w - \frac{2b}{w} l_b, \\
&\dot{\mathbf{a}} = -b l_{\mathbf{v}} + 2 l_{\mathbf{a}}, \quad \dot{b} = -\frac{b^2+1}{w^2} \mathbf{a}^{\mathrm{T}} l_{\mathbf{v}} - \frac{2b}{w} l_w + \frac{2(b^2+1)}{w^2} l_b.
\end{aligned}$$

Then, we fix $(\mathbf{v},\, w)$ to its best approximation $(\mathbf{v}^*, w^*)$. We examined the evolution of parameter $(\mathbf{a},\, b)$. For $\xi^* = (\mathbf{v}^*, w^*, \mathbf{0}, b)$, we defined

$$H(\mathbf{v}^*, w^*) := \frac{1}{4} w^* \left\langle e(y, \mathbf{x}, \xi) \frac{\partial^2 \phi(\mathbf{x}, \mathbf{v})}{\partial \mathbf{v} \partial \mathbf{v}^{\mathrm{T}}} \right\rangle \Bigg|_{\xi=\xi^*}.$$

For loss function $l(y, \mathbf{x}, \xi)$, $\left\langle \frac{\partial^2 l(y,\mathbf{x},\xi)}{\partial \xi \partial \xi^{\mathrm{T}}} \right\rangle \Big|_{\xi=\xi^*} = (1-b^2)H(\mathbf{v}^*, w^*)$ holds.

Theorem 1 (Stability of learning near singular regions) *For stability of learning near singular regions, it holds as follows: When the true density function is in a singular region, the entire critical line of R_0 is stable. When the true density function is not in a singular region, the stability of the entire critical line of R_0 is divided into the following three cases according to the eigenvalue of $H(\mathbf{v}^*, w^*)$[5].*

(1) both positive and negative eigenvalues: all points on the critical line of R_0 are unstable.

(2) negative definite: the part $b^2 < 1$ is stable, whereas the part $b^2 < 1$ is unstable in R_0.

(3) positive definite: the part $b^2 < 1$ is stable, whereas the part $b^2 > 1$ is unstable in R_0.

We assume $\tilde{\xi} = (\mathbf{v}^*, w^*, \mathbf{a}, b)$. The gradient of loss function $l(\tilde{\xi})$ holds as follows[3]:

$$\begin{aligned}
&l_{\mathbf{v}}(\tilde{\xi}) = \frac{1}{8} w^*(1-z^2)Q(\mathbf{v}^*, \mathbf{a}) + O(\mathbf{a}^3), \;\; l_w(\tilde{\xi}) = \frac{1}{2}\frac{1-z^2}{w^*}\mathbf{a}^{\mathrm{T}} H(\mathbf{v}^*, w^*)\mathbf{a} + O(\mathbf{a}^3),\\
&l_{\mathbf{a}}(\tilde{\xi}) = (1-z^2)H(\mathbf{v}^*, w^*)\mathbf{a} + \frac{1}{24} w^* z(1-z^2) \left\langle e(y, \mathbf{x}, \xi) \frac{\partial D(x, \mathbf{v}, \mathbf{a})}{\partial \mathbf{a}} \right\rangle \Bigg|_{\xi=\tilde{\xi}} + O(\mathbf{a}^3),\\
&l_b(\tilde{\xi}) = -b\, \mathbf{a}^{\mathrm{T}} H(\mathbf{v}^*, w^*)\mathbf{a} + O(\mathbf{a}^3).
\end{aligned}$$

Note that $l_{\mathbf{a}}(\tilde{\xi})$ is of order $O(\mathbf{a})$ and $l_b(\tilde{\xi})$, $l_{\mathbf{v}}(\tilde{\xi})$, $l_w(\tilde{\xi})$ is of order $O(\mathbf{a}^2)$. Neglecting higher terms in the above equations and taking into account the fact that $\mathbf{a} \approx \mathbf{0}$, the learning equation near R_0 holds as follows[3]:

$$\dot{\mathbf{a}} = 2(1-b^2)H(\mathbf{v}^*, w^*)\mathbf{a}, \quad \dot{b} = -\frac{b(1-b^2)}{w^{*2}}\mathbf{a}^{\mathrm{T}} H(\mathbf{v}^*, w^*)\mathbf{a} - \frac{2b(b^2+1)}{w^{*2}}\mathbf{a}^{\mathrm{T}} H(\mathbf{v}^*, w^*)\mathbf{a}.$$

Theorem 2 (Dynamics of learning near singular regions) *An energy function $h(\mathbf{a}) := \frac{1}{2}\mathbf{a}^{\mathrm{T}}\mathbf{a}$ of the dynamics of learning near singular regions, it holds as follows[3]:*

(1) In the neighborhood of R_0, we obtain the equation $\dot{h} = \mathbf{a}^{\mathrm{T}}\dot{\mathbf{a}} = \frac{2w^{*2}(b^2-1)}{b(b^2+3)}\dot{b}$ and the dynamics of the learning equations are given by

$$h(\mathbf{a}) = \frac{2w^{*2}}{3} \log \frac{(b^2+3)^2}{|b|} + C.$$

(2) In the neighborhood of $R_0 \cap R_1$, we obtain the equation $\dot{h} = \frac{w^{*2}(b^2-1)}{b(b^2+1)}\dot{b}$ and the dynamics of the learning equations are given by

$$h(\mathbf{a}) = w^{*2} \log \left(|b| + \frac{1}{|b|} \right) + C.$$

Definition 7 (Classification of dynamics of learning near singular regions) *The dynamics of learning near a singularity is classified into following five patterns by changing an initial value of the statistical model[6].*

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

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

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

(4) Near elimination singularity: When the parameters of the statistical model are near the elimination singularity in the training, the learning process is significantly affected by elimination singularity.

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

3 Construction of a neural network as the statistical model using Mathematica.

Using Mathematica, variables $F1,\ F2$, constants $elem0,\ elem1,\ elem2,\ elem3$ calculate as follows:

$F1[a\_] := NetInsertSharedArrays[NetChain[LinearLayer[1, "Weights" - > a, "Biases" - > None]], "Linear1"]$,
$F2[b\_] := NetInsertSharedArrays[NetChain[LinearLayer[1, "Weights" - > b, "Biases" - > None]], "Linear2"]$,
$elem0 := ElementwiseLayer[\# * (1/2)\&],\ elem1 := ElementwiseLayer[\# * (-1)\&]$,
$elem2[v\_] := ElementwiseLayer[\# * (v)\&],\ elem3[w\_] := ElementwiseLayer[\# * (w)\&]$.

First, to express the condition $w_{11}x = \left(v + \frac{1}{2}(b-1)a\right) x$, we input the following:

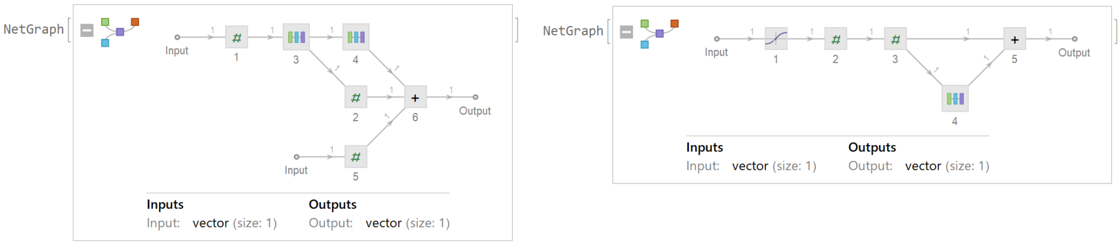
$$net11[a\_, b\_, v\_] := NetGraph[elem0, elem1, F1[a], F2[b], elem2[v], TotalLayer[],$$
$$NetPort["Input"] - > 1, 1 - > 3 - > 4, 3 - > 2, 4, 2, 5 - > 6]$$

and $net11$ output on the left-hand side of Figure 1.

To express the condition $w_{31}\tanh(x) = \frac{1}{2}w(b+1)\tanh(x)$, we input the following:

$$net12[a\_, b\_, w\_] := NetGraph[Tanh, elem0, elem3[w], F2[b], TotalLayer[],$$
$$NetPort["Input"] - > 1, 1 - > 2 - > 3 - > 4, 3, 4 - > 5]$$

and $net12$ output on the light-hand side of Figure 1.



Inputs
Input: vector (size: 1)

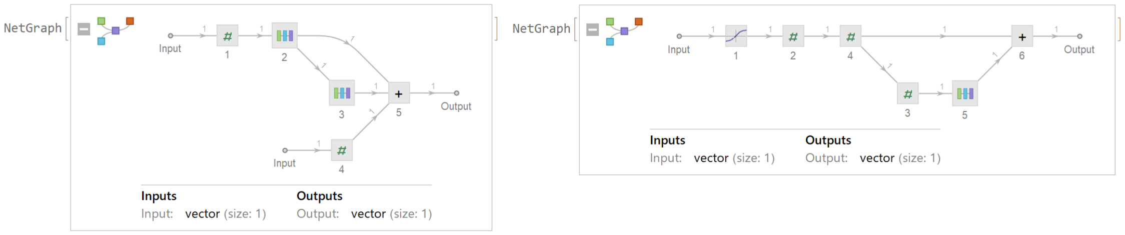
Outputs
Output: vector (size: 1)

Inputs
Input: vector (size: 1)

Outputs
Output: vector (size: 1)

Figure 1: $net11,\ net12$

Similarly to express the condition $w_{12}x = \left(v+\frac{1}{2}(b+1)a\right)x$, $w_{32}\tanh(x) = \frac{1}{2}w(-b+1)\tanh(x)$, similarly we input in the same way. $net21$, $net22$ output on the Figure 2.



Inputs
Input: vector (size: 1)

Outputs
Output: vector (size: 1)

Inputs
Input: vector (size: 1)

Outputs
Output: vector (size: 1)

Figure 2: $net21$, $net22$

Next to express the condition $w_{31}\tanh(w_{11}x) = \frac{1}{2}w(b+1)\tanh[\left(v+\frac{1}{2}(b-1)a\right)x]$, we input the following:

$$net1[a\_, b\_, v\_, w\_] := NetGraph[\mathrm{net}11[a,b,v], \mathrm{net}12[a,b,w], NetPort[''Input'']- > 1, 1- > 2]$$

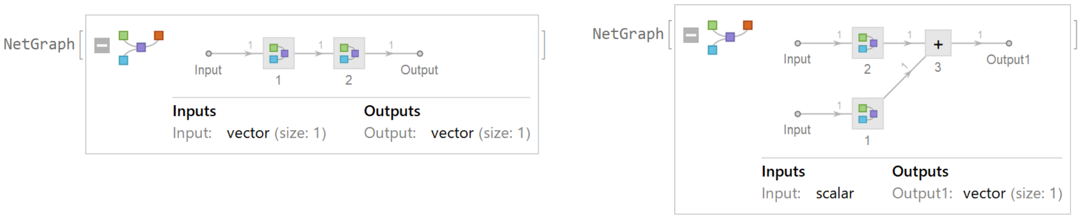
and $net1$ output on the left-hand side of Figure 3.

Similarly, to express the condition $w_{32}\tanh(w_{12}x) = \frac{1}{2}w(-b+1)\tanh[\left(v+\frac{1}{2}(b+1)a\right)x]$, we defined $net2$.

Finally, to express the condition $w_{31}\tanh(w_{11}x)+w_{32}\tanh(w_{12}x)$, we input the following:

$$parameterNet[a\_, b\_, v\_, w\_] := NetGraph[net1[a,b,v,w], net2[a,b,v,w], TotalLayer[],$$
$$NetPort[''Input'']- > 1, NetPort[''Input'']- > 2{,}1, 2- > 3- > NetPort[''Output1''], ''Input''- > enc]$$

and $parameterNet$ output on the light-hand side of Figure 3.



Inputs
Input: vector (size: 1)

Outputs
Output: vector (size: 1)

Inputs
Input: scalar

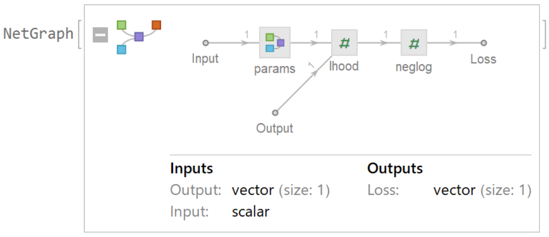
Outputs
Output1: vector (size: 1)

Figure 3: $net1$, $parameterNet$

Let us define the loss function as a log density ratio function. We input the following:

$$gaussianLikelihood[y\_, \mu\_] := PDF[NormalDistribution[\mu, 1], y]$$
$$trainingNet[a\_, b\_, v\_, w\_] := NetGraph[<|''params''- > parameterNet[a,b,v,w], ''lhood''- >$$
$$ThreadingLayer[gaussianLikelihood], ''neglog''- > ElementwiseLayer[-Log[\#]\&]|>,$$
$$NetPort[''Output''], NetPort[''params'', ''Output1'']- > ''lhood'', ''lhood''- > ''neglog''- > NetPort[''Loss'']]$$

and $trainingNet$ output on the light side of Figure 4.



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

Outputs
Loss: vector (size: 1)

Figure 4: $trainingNet(log\,density\,ratio)$

For training data and test data, we input the following:

$$G[a\_, b\_] := Mean[trainingNet[a, b, v0, w0][< |"Input" - > dataX, "Output" - > enc[dataY]| >]]$$
$$H[a\_, b\_] := Mean[trainingNet[a, b, v0, w0][< |"Input" - > testX, "Output" - > enc[testY]| >]]$$

and defined training loss function G and validation loss function H.

4 Dynamics of learning near singular regions

4.1 Framework of dynamics of learning

Example 1 (Training data, true density function) *For input X on $-3 \leq x \leq 3$ and noise Z of $\sigma = 0.05$, let the training data (are listed in Appendix.) be*

$$0.25\tanh(0.2x) + 0.25\tanh(0.4x) + Z,$$

and the true density function be

$$q(y|x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{|y - (0.25\tanh(0.2x) + 0.25\tanh(0.4x))\,|^2}{2\sigma^2}\right).$$

For $a = 0.2$, $b = 0$, $v = 0.3$, $w = 0.5$, we consider that the dynamics of learning evolving under the influence of a critical line classified into five cases by changing the initial values of the statistical model for the case in which the true distribution near the singular regions is realizable by the statistical model. Let us define the loss function as the log density ratio function, and input the following:

$$results1[a\_, b\_] := NetTrain[trainingNet[a, b, v, w], < |"Input" - > dataX, "Output" - > enc[dataY]| >,$$
$$\{"Weights", "TrainedNet", "RoundLossList"\}, LossFunction - > "Loss", Method - >$$
$$"ADAM", "LearningRate" - > 0.1, BatchSize - > 30, MaxTrainingRounds - > \{\quad\},$$
$$TrainingProgressFunction - > appendToLog]$$

In addition, the neural network was trained.

4.2 Dynamics of overlap singularity and cross elimination singularity

Let the initial values of the statistical model be $a = 0.15$, $b = -2.0$, -1.8, -1.5, -1.3, $v = 0.3$, $w = 0.5$. The neural network was trained 140 times. We construct an array of parameters of a, b under the influence of the critical line. The evolutions of parameters of a and the evolutions of parameters of b are shown on the left-hand and middle-hand sides respectively, of Figure 5, and the evolution of parameters of a, b is shown on the light-hand side of Figure 5.

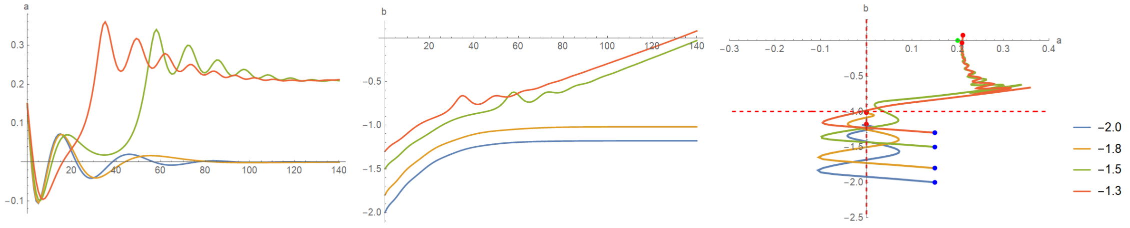


Figure 5: Evolution of parameters of a, b ($a = 0.15$, $b = -2.0$, -1.8, -1.5, -1.3)

The neural network was trained 140 times. We construct an array of the training loss, and the evolution of the training loss and the dynamics of learning of the training loss surface are shown on the left-hand and right-hand sides, respectively, of Figure 6.

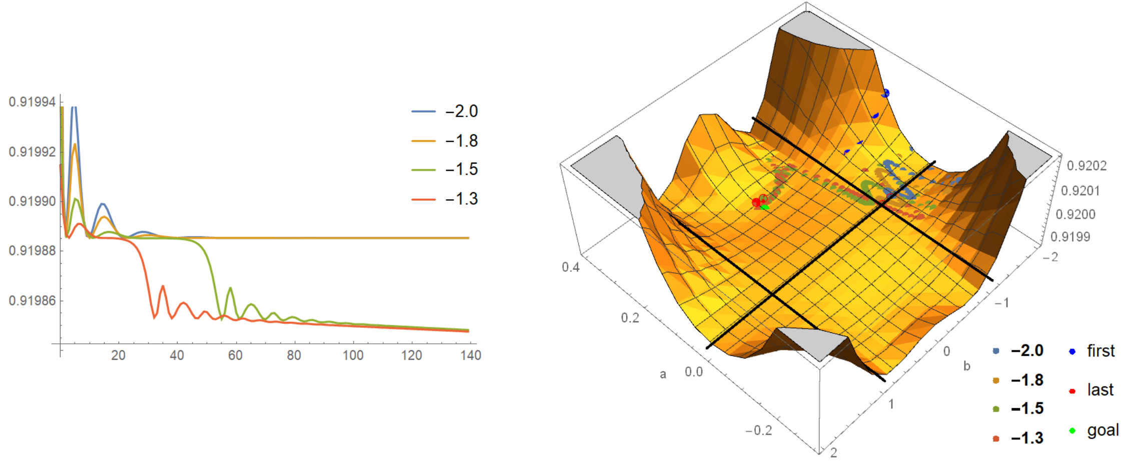


Figure 6: Evolution of the training loss and the dynamics of the training loss surface ($a = 0.15$, $b = -2.0,\ -1.8,\ -1.5,\ -1.3$)

We generalize the parameter of $b\,(-2.2 \leq b \leq 2.2)$. The evolution of the parameters of a, b and the dynamics of learning of the training loss surface are shown on the left-hand and right-hand sides, respectively, of Figure 7.

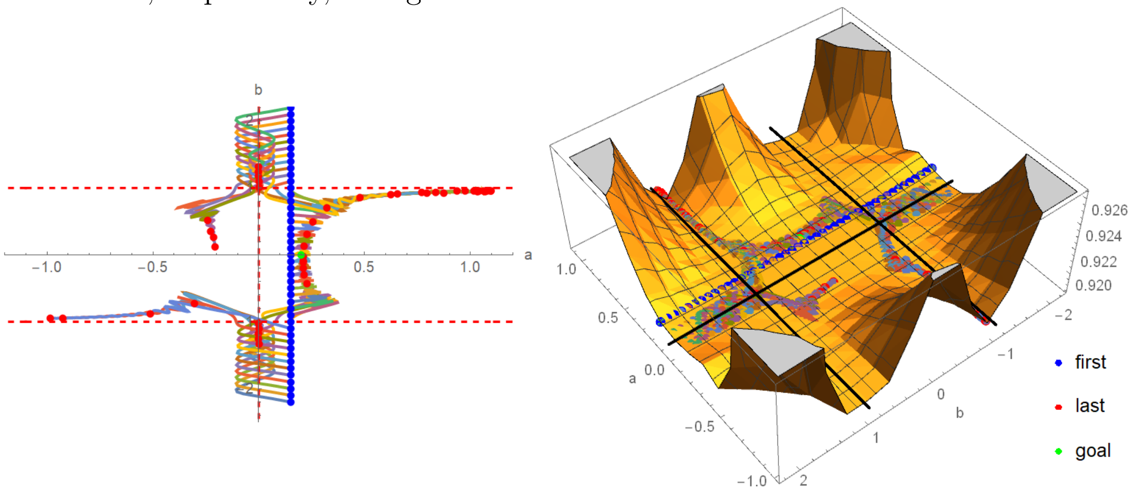


Figure 7: Evolution of the parameters of a, b and the dynamics of the training loss surface ($-2.2 \leq b \leq 2.2$)

Result 1 *(1) We find that plateau phenomena were observed on critical line* $a = 0$ *and that the dynamics of learning do not reach the true distribution in case* $b = -2.0,\ -1.8$.

(2) We find that plateau phenomena were observed when crossing critical line $b = -1$ *and that the dynamics of learning reach the true distribution in case* $b = -1.5\ -1.3$.

(3) As the parameter b evolves to 0, the dynamics of learning change from overlap singularity to cross elimination singularity and from cross elimination singularity to fast convergence.

4.3 Dynamics of near elimination singularity and output weight 0

Let the initial values of the statistical model be $a = 0.5,\ 0.6,\ 0.7,\ 1.2, b = 0.75, v = 0.3, w = 0.5$. The neural network was trained 100 times. We construct an array of parameters of a, b under the influence of critical line. The evolutions of parameters of a and the evolutions of parameters of b are shown on the left-hand and middle-hand sides respectively, of Figure 8, and the evolution of parameters of a, b is shown on the light-hand side of Figure 8.

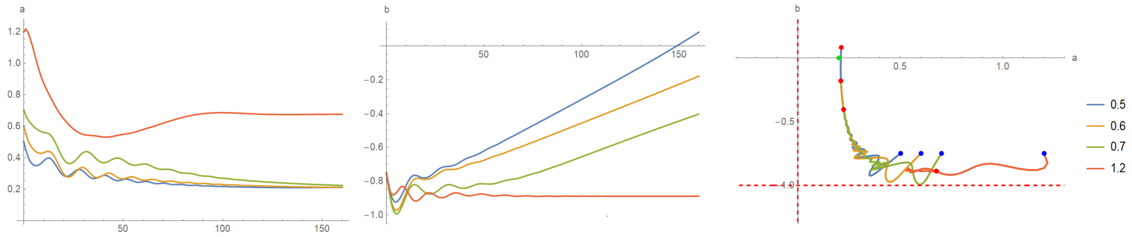


Figure 8: Evolutions of parameters of a, b ($a = 0.5,\ 0.6,\ 0.7,\ 1.2,\ b = 0.75$)

The neural network was trained 100 times. We construct an array of the training loss, and the evolution of the training loss and the dynamics of learning of the training loss surface are shown on the left-hand and right-hand sides, respectively, of Figure 9.

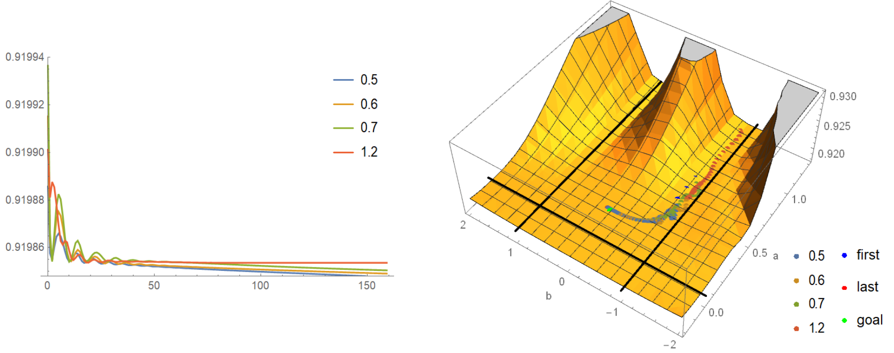


Figure 9: Evolution of the training loss and the dynamics of the training loss surface ($a = 0.5,\ 0.6,\ 0.7,\ 1.2,\ b = 0.75$)

We generalize the parameter of $a\,(0 \leq a \leq 2.2)$. The evolution of the parameters of a, b and the dynamics of learning of the training loss surface are shown on the left-hand and right-hand sides, respectively, of Figure 10.

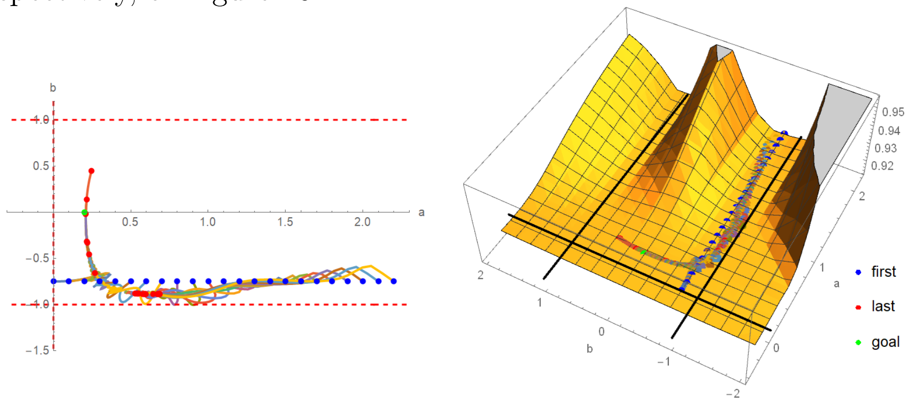


Figure 10: Evolution of the parameters of a, b and the dynamics of the training loss surface ($0 \leq a \leq 2.2$)

Result 2 *(1) We find that plateau phenomena were observed on critical line $b = -1$ and that the dynamics of learning do not reach the true distribution in case $a = 1.2$.*

(2) We find that plateau phenomena were observed approaching critical line $b = -1$ and that the dynamics of learning reach the true distribution in case $a = 0.6, 0.7$. Moreover, we find that the dynamics of learning reach the true distribution more quickly in case $a = 0.5$.

(3) As the parameter of a evolves to 0, the dynamics of learning change from output weight 0 to near elimination singularity and from near elimination singularity to fast convergence.

5 Conclusion

Firstly we constructed the neural network as the statistical model using Mathmatica. Secondly, we observed plateau phenomena near singular regions by examining the evolution of the parameter and the dynamics of learning on the training loss surface. Finally we investigated that the type of dynamics of learning changes when the overlap and the elimination singularity is approached from a distance by changing the initial values of the statistical model.

The purpose of our research is not to study the current state of the art in neural networks but to make some concepts in phenomena in neural networks correspond (explain) to learning in educational activities. Specifically, it is to clarify what state the phenomenon of plateau and over-fitting and over-generalization are in educational activities.

The results can also be the foundation to investigate the singular learning dynamics in educational activities.

References

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

[2] S. Watanabe, “ Mathematical Theory of Bayesian Statistics, ” CRC Press, 2018.

[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] F. Cousseau, T. Ozeki, and S. Amari, “ Dynamics of Learning in Multilayer Perceptrons Near Singularities,” IEEE Transactions on Neural Networks, vol. 19, no. 8, pp. 1313–1328, 2008.

[5] K. Fukumizu and S. Amari, “ Local minima and plateaus in hierarchical structures of multilayer perceptrons, ” Neural Networks, vol. 13, no. 3, pp. 317–327, 2000.

[6] W. Guo, H. Wei , Y. Ong, J. R. Hervas, J. Zhao, H. Wang, K. Zhang, “Numerical Analysis near Singularities in RBF Networks, ” Journal of Machine Learning Research, vol. 19, no. 1, pp. 1-39, 2018.

6 Appendix

6.1 Training data of Example1

For input x_s, and output y_s as follows:

$$\begin{aligned}
&x_s = \{2.467685732795669, 1.6313896711002975, 1.7039693114471142, -2.353539095169551,\\
&2.5106926463104458, -2.9742536653063, 1.4778884503387921, -1.7619315572659175,\\
&0.8575206146347014, -1.9522402751318726, -2.9186556422433796, 2.3433244821789305,\\
&-2.3174593747595598, 0.24745360478229195, -0.43473282858294837, 2.0777962243403962,\\
&-0.7489587340884398, 0.40283200240701333, 1.4393667305075848, 2.6884952319559243,\\
&0.4233060018195829, 1.3133371734415373, -1.8687861826912897, 2.641499809476027,\\
&1.3536619131864676, 1.4261447937286373, -1.5373889449365947, 2.5833410435168336,\\
&-0.7634883775841974, -1.418229957030034\},
\end{aligned}$$

$$\begin{aligned}
&y_s = \{0.25047549494898375, 0.14195709642758433, 0.2416763776071971, -0.34055590890961035,\\
&0.3082658314034902, -0.4292549244954509, 0.15038776701404105, -0.23410034295044008,\\
&0.1674469014375939, -0.26548937037643955, -0.3321460817933551, 0.2720167181157782,\\
&-0.2892455062624837, 0.0520546848151971, -0.0009290519327547, 0.2940081059525326,\\
&-0.14421683321295234, 0.08562704853302514, 0.25724997978192643, 0.2668005655536598,\\
&0.043918697553646746, 0.19753643159405437, -0.2627853499983649, 0.25989101875041354,\\
&0.1395086144673041, 0.17062611740258554, -0.18466386529707135, 0.3690548490941195,\\
&-0.16241114605952112, -0.14051890248769974\},
\end{aligned}$$

Then, we defined training data as follows:

$$\begin{aligned}
&\{2.46769 \to 0.250475, 1.63139 \to 0.141957, 1.70397 \to 0.241676, -2.35354 \to -0.340556,\\
&2.51069 \to 0.308266, -2.97425 \to -0.429255, 1.47789 \to 0.150388, -1.76193 \to -0.2341,\\
&0.857521 \to 0.167447, -1.95224 \to -0.265489, -2.91866 \to -0.332146, 2.34332 \to 0.272017,\\
&-2.31746 \to -0.289246, 0.247454 \to 0.0520547, -0.434733 \to -0.000929052,\\
&2.0778 \to 0.294008, -0.748959 \to -0.144217, 0.402832 \to 0.085627, 1.43937 \to 0.25725,\\
&2.6885 \to 0.266801, 0.423306 \to 0.0439187, 1.31334 \to 0.197536, -1.86879 \to -0.262785,\\
&2.6415 \to 0.259891, 1.35366 \to 0.139509, 1.42614 \to 0.170626, -1.53739 \to -0.184664,\\
&2.58334 \to 0.369055, -0.763488 \to -0.162411, -1.41823 \to -0.140519\}.
\end{aligned}$$