Analytical form of Fisher information matrix of bipoloar-activation-function-based multilayer perceptrons

Weili Guo*†, Liping Xie†, Zhenyong Fu*, Jianhui Guo*, Guochen Pang‡, Jian Yang*
* *PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and Jiangsu Key Lab of Image and Video Understanding for Social Security, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, P.R. China*
† *Key Laboratory of Measurement and Control of CSE, Ministry of Education, School of Automation, Southeast University, Nanjing, P.R. China*
‡ *School of Automation and Electrical Engineering, Linyi University, Linyi, P.R. China*
wlguo@njust.edu.cn, lpxie@seu.edu.cn, z.fu@njust.edu.cn, guojianhui@njust.edu.cn, guochenpang@163.com, csjyang@njust.edu.cn

***Abstract*—For the widely used multilayer perceptrons (MLPs), the existed singularities in the parameter space have seriously affected the learning dynamics of MLPs, which cause several singular learning behaviors. Since the Fisher information matrix (FIM) plays a significant role in analyzing the singular learning dyanmics of MLPs, it is very worthy to obtain the analytical form of FIM to do further investigation. In this paper, by choosing the bipolar error function as the activation function, the analytical form of FIM are obtained, where the validity of the obtained results are verified by taking three experiments.**

***Index Terms*—Fisher information matrix, multilayer perceptrons, singularity, bipolar error function, feedforward neural networks**

I. Introduction

The multilayer perceptron (MLP) is a typical class of feedforward neural networks, which has been widely used in many fields, such as pattern recognition, intelligent control and artificial intelligence, etc [1–3]. When using MLPs in various applications, researchers found some different learning dynamics during the learning process in comparison with the regular learning machines, which are called singular learning dynamics. For example, there are many local minima in the parameter space, the learning process may suddenly become very slow and plateau phenomenon can often be observed (an example is shown in Fig. 1) [4–6].

Many researchers have investigated the singular learning dynamics and obtain the analysis results that the existed singularities in the parameter space mainly cause the singular behaviors [7–9]. These singularities are the subspaces of parameter space where the Fisher information matrix (FIM) is singular [10]. As the FIM degenerates on the singularities, the classic paradigm of the Cramer-Rao theorem in the regular models does not hold [11].

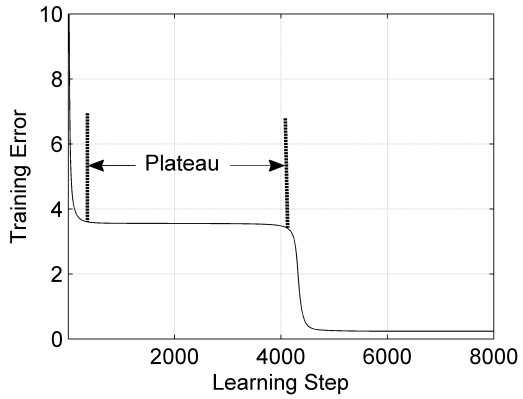


Fig. 1. Plateau phenomenon occurred in the learning process of MLPs

In view of the serious influence of singularities on the learning dynamics of feedforward neural networks, many researchers have investigated the learning dynamics near singularities in types of feedforward neural networks, such as the multilayer perceptrons [12] radial basis function (RBF) networks[13], and Gaussian mixture model[14] etc. For these neural networks, different cases have been carefully analyzed, including toy model case [15], regular case [16], and unrealizable case [17]. Further, researcher also investigate how to overcome the influence of singularities in feedforward neural networks. Due to the irreversibility of FIM on singularities, instead of the gradient descent direction, Riemann gradient (natural gradient) descent direction becomes the steepest descent direction [18]. Thus natural gradient descent method has been proposed to accelerate the learning process [19], where the inverse of FIM is added to the modified formulation of standard gradient descent algorithm. Then many modified

This work was supported by the National Science Fund of China under Grant Nos. 61906092, 61876085, U1713208, 61472187 and 61603190, Natural Science Foundation of Jiangsu Province of China under Grant No. BK20190441, the 973 Program No. 2014CB349303, and Program for Changjiang Scholars.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

natural gradient algorithms are designed to decrease the cost of computation in obtaining the inverse of FIM [20–23].

Given that FIM plays a key role in investigating the singular learning dynamics, it is very meaningful to obtain the analytical form of FIM. Based on the obtained analytical form, we can design modified natural gradient descent algorithms with more efficiency and better performances further. For the bipolar-activation-function-based MLPs, hyperbolic tangent function $\tanh(x) = \dfrac{e^x - e^{-x}}{e^x + e^{-x}}$ is the most used activation function. However, as hyperbolic tangent function is non-integrable, we can not obtain the analytical form of FIM for hyperbolic-tangent-function-based case. In order to overcome this problem, we choose the bipolar error function $\phi(x) = \sqrt{\dfrac{2}{\pi}} \displaystyle\int_0^x \exp\left(-\frac{1}{2}t^2\right) \mathrm{d}t$ as the activation function [16], which is also of the sigmoid function and is integral, then we can get the analytical form of FIM in this paper .

The rest of this paper is organized as follows. In section 2, we give the analytical form of FIM. The efficiency of the obtained results are verified by taking simulation experiments in section 3. Section 4 states conclusions and discussions.

II. Analytical Form of Fisher Information Matrix

In this section, we first introduce the learning paradigm, and then the analytical form of FIM can be obtained.

The multilayer perceptrons with k hidden nodes which receive input $\boldsymbol{x}$ can be represented as follows:

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{i=1}^{k} w_i \phi(\boldsymbol{x}, \boldsymbol{J}_i), \tag{1}$$

where $\boldsymbol{J}_i$ is the weight from the input layer to the hidden node i and w_i is the weight from the hidden node i to the output layer. $\phi(\cdot)$ is the activation function, and $\phi(\boldsymbol{x}, \boldsymbol{J}_i) = \phi(\boldsymbol{J}_i^T \boldsymbol{x})$ is the output of the i-th hidden neuron. Thus $\boldsymbol{\theta} = \{\boldsymbol{J}_1, \cdots, \boldsymbol{J}_k, w_1, \cdots, w_k\}$ represents all the parameters of the model. Here the bipolar error function is chosen as the activation function, namely $\phi(\boldsymbol{x}, \boldsymbol{J}_i) = \sqrt{\dfrac{2}{\pi}} \displaystyle\int_0^{\boldsymbol{J}_i^T \boldsymbol{x}} \exp\left(-\frac{1}{2}t^2\right) \mathrm{d}t$.

For the regression problem, we have a number of observed data $(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_t, y_t)$, which are generated by an unknown teacher function:

$$y = f_0(\boldsymbol{x}) + \varepsilon, \tag{2}$$

where the additive noise ε subjects to Gaussian distribution with zero mean and variance σ_0^2, and the observed data are used to training MLPs in order to approximate the teacher function.

Without loss of generality, the training input is assumed to be subject to a standard Gaussian distribution, i.e.:

$$q(\boldsymbol{x}) = (\sqrt{2\pi})^{-n} \exp\left(-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{x}\right). \tag{3}$$

The loss function is defined as:

$$l(y, \boldsymbol{x}, \boldsymbol{\theta}) = \frac{1}{2}(y - f(\boldsymbol{x}, \boldsymbol{\theta}))^2, \tag{4}$$

and we use the gradient descent method to minimize the above loss:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \frac{\partial l(y_t, \boldsymbol{x}_t, \boldsymbol{\theta}_t)}{\partial \boldsymbol{\theta}_t}, \tag{5}$$

where η represents the learning rate.

For the bipolar-activation-function-based MLPs (1), there exist at least two types of singularities in the parameter space as follows [16]:

(1) Opposite singularity: $\mathcal{R}_1 = \{\boldsymbol{\theta} | \boldsymbol{J}_i = -\boldsymbol{J}_j\}$,

(2) Elimination singularity: $\mathcal{R}_2 = \{\boldsymbol{\theta} | w_i = 0\}$.

The FIM can be defined in the following equation [7]:

$$\boldsymbol{F}(\boldsymbol{\theta}) = \left\langle \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right\rangle. \tag{6}$$

$\langle \cdot \rangle$ denotes the expectation with respect to the teacher distribution, which is given by:

$$p_0(y, \boldsymbol{x}) = q(\boldsymbol{x}) \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{(y - f_0(\boldsymbol{x}))^2}{2\sigma_0^2}\right). \tag{7}$$

Here we obtain the explicit expressions of $\left\langle \dfrac{\partial \phi(\boldsymbol{x}, \boldsymbol{J}_i)}{\partial \boldsymbol{J}_i} \dfrac{\partial \phi(\boldsymbol{x}, \boldsymbol{J}_j)}{\partial \boldsymbol{J}_j^T} \right\rangle$, $\left\langle \dfrac{\partial \phi(\boldsymbol{x}, \boldsymbol{J}_i)}{\partial \boldsymbol{J}_i} \phi(\boldsymbol{x}, \boldsymbol{J}_j) \right\rangle$, and $\langle \phi(\boldsymbol{x}, \boldsymbol{J}_i) \phi(\boldsymbol{x}, \boldsymbol{J}_j) \rangle$ at first, which are the fundamental to get the analytical form of FIM. For simplicity, we denote: $\boldsymbol{Q}_1(\boldsymbol{J}_i, \boldsymbol{J}_j) = \left\langle \dfrac{\partial \phi(\boldsymbol{x}, \boldsymbol{J}_i)}{\partial \boldsymbol{J}_i} \dfrac{\partial \phi(\boldsymbol{x}, \boldsymbol{J}_j)}{\partial \boldsymbol{J}_j^T} \right\rangle$, $\boldsymbol{Q}_2(\boldsymbol{J}_i, \boldsymbol{J}_j) = \left\langle \dfrac{\partial \phi(\boldsymbol{x}, \boldsymbol{J}_i)}{\partial \boldsymbol{J}_i} \phi(\boldsymbol{x}, \boldsymbol{J}_j) \right\rangle$, and $Q_3(\boldsymbol{J}_i, \boldsymbol{J}_j) = \langle \phi(\boldsymbol{x}, \boldsymbol{J}_i) \phi(\boldsymbol{x}, \boldsymbol{J}_j) \rangle$.

In the following lemma, we give the explicit expressions of $\boldsymbol{Q}_1(\boldsymbol{J}_i, \boldsymbol{J}_j)$, $\boldsymbol{Q}_2(\boldsymbol{J}_i, \boldsymbol{J}_j)$ and $Q_3(\boldsymbol{J}_i, \boldsymbol{J}_j)$.

Lemma 1: The analytical forms of $\boldsymbol{Q}_1(\boldsymbol{J}_i, \boldsymbol{J}_j)$, $\boldsymbol{Q}_2(\boldsymbol{J}_i, \boldsymbol{J}_j)$ and $Q_3(\boldsymbol{J}_i, \boldsymbol{J}_j)$ are given as follows:

$$\boldsymbol{Q}_1(\boldsymbol{J}_i, \boldsymbol{J}_j) = \frac{2}{\pi} \sqrt{\det(\boldsymbol{B}^{-1}(\boldsymbol{J}_i, \boldsymbol{J}_j))} \boldsymbol{B}^{-1}(\boldsymbol{J_i}, \boldsymbol{J_j}), \tag{8}$$

$$\boldsymbol{Q}_2(\boldsymbol{J}_i, \boldsymbol{J}_j) = \frac{2}{\pi} \sqrt{\det(\boldsymbol{B}^{-1}(\boldsymbol{J}_i, \boldsymbol{J}_j))} \boldsymbol{A}^{-1}(\boldsymbol{J}_i) \boldsymbol{J}_j, \tag{9}$$

$$Q_3(\boldsymbol{J}_i, \boldsymbol{J}_j) = \frac{2}{\pi} \arcsin \frac{\boldsymbol{J}_i^T \boldsymbol{J}_j}{\sqrt{1 + \boldsymbol{J}_i^T \boldsymbol{J}_i} \sqrt{1 + \boldsymbol{J}_j^T \boldsymbol{J}_j}}, \tag{10}$$

where:

$$\boldsymbol{A}(\boldsymbol{J}_i) = \mathbf{I}_n + \boldsymbol{J}_i \boldsymbol{J}_i^T, \tag{11}$$

$$\boldsymbol{B}(\boldsymbol{J}_i, \boldsymbol{J}_j) = \boldsymbol{A}(\boldsymbol{J}_i) + \boldsymbol{J}_j \boldsymbol{J}_j^T = \mathbf{I}_n + \boldsymbol{J}_i \boldsymbol{J}_i^T + \boldsymbol{J}_j \boldsymbol{J}_j^T, \tag{12}$$

$$\boldsymbol{A}^{-1}(\boldsymbol{J}_i) = \left(\mathbf{I}_n + \boldsymbol{J}_i \boldsymbol{J}_i^T\right)^{-1} = \mathbf{I}_n - \frac{\boldsymbol{J}_i \boldsymbol{J}_i^T}{1 + \|\boldsymbol{J}_i\|^2}, \tag{13}$$

$$\mathbf{B}^{-1}(\boldsymbol{J}_i, \boldsymbol{J}_j) = \boldsymbol{A}^{-1}(\boldsymbol{J}_i) - \frac{\boldsymbol{A}^{-1}(\boldsymbol{J}_i) \boldsymbol{J}_j \boldsymbol{J}_j^T \boldsymbol{A}^{-1}(\boldsymbol{J}_i)}{1 + \boldsymbol{J}_j^T \boldsymbol{A}^{-1}(\boldsymbol{J}_i) \boldsymbol{J}_j}, \tag{14}$$

$$\det(\boldsymbol{B}^{-1}(\boldsymbol{J}_i, \boldsymbol{J}_j)) = \frac{1}{(1 + \|\boldsymbol{J}_i\|^2)(1 + \|\boldsymbol{J}_j\|^2) - (\boldsymbol{J}_i^T \boldsymbol{J}_j)^2}, \tag{15}$$

$\mathbf{I}_n$ is the compatible identity matrix.

Proof: The calculation process is shown in Appendix A. □

Then we can obtain the analytical form of the FIM which is shown in Theorem 1.

Theorem 1: The analytical form of the Fisher information matrix $\boldsymbol{F}(\boldsymbol{\theta})$ is given in equation (16).

Proof: By substituting $f(\boldsymbol{x},\boldsymbol{\theta}) = \sum_{i=1}^{k} w_i \phi(\boldsymbol{x},\boldsymbol{J}_i)$ into equation (6), we have:

$$\boldsymbol{F}(\boldsymbol{\theta}) = \left\langle \frac{\partial f(\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial f(\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right\rangle = \begin{bmatrix} F_{11} & \cdots & F_{1k} & F_{1(k+1)} & \cdots & F_{1(2k)} \\ \vdots & & \vdots & \vdots & & \vdots \\ F_{k1} & \cdots & F_{kk} & F_{k(k+1)} & \cdots & F_{k(2k)} \\ F_{(k+1)1} & \cdots & F_{(k+1)k} & F_{(k+1)(k+1)} & \cdots & F_{(k+1)(2k)} \\ \vdots & & \vdots & \vdots & & \vdots \\ F_{(2k)1} & \cdots & F_{(2k)k} & F_{(2k)(k+1)} & \cdots & F_{(2k)(2k)} \end{bmatrix} \tag{17}$$

where:

$$\begin{aligned} F_{ij} &= \left\langle \frac{\partial f(\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{J}_i} \frac{\partial f(\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{J}_j^T} \right\rangle \\ &= w_i w_j \left\langle \frac{\partial \phi(\boldsymbol{x},\boldsymbol{J}_i)}{\partial \boldsymbol{J}_i} \frac{\partial \phi(\boldsymbol{x},\boldsymbol{J}_j)}{\partial \boldsymbol{J}_j^T} \right\rangle \\ &= w_i w_j \boldsymbol{Q}_1(\boldsymbol{J}_i,\boldsymbol{J}_j), \quad \text{for } 1 \le i \le k, \quad 1 \le j \le k, \end{aligned} \tag{18}$$

$$\begin{aligned} F_{ij} &= \left\langle \frac{\partial f(\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{J}_i} \frac{\partial f(\boldsymbol{x},\boldsymbol{\theta})}{\partial w_{j-k}} \right\rangle \\ &= w_i \left\langle \frac{\partial \phi(\boldsymbol{x},\boldsymbol{J}_i)}{\partial \boldsymbol{J}_i} \phi(\boldsymbol{x},\boldsymbol{J}_{j-k}) \right\rangle \\ &= w_i \boldsymbol{Q}_2(\boldsymbol{J}_i,\boldsymbol{J}_{j-k}), \quad \text{for } 1 \le i \le k, \ k+1 \le j \le 2k, \end{aligned} \tag{19}$$

$$\begin{aligned} F_{ij} &= \left\langle \frac{\partial f(\boldsymbol{x},\boldsymbol{\theta})}{\partial w_{i-k}} \frac{\partial f(\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{J}_j^T} \right\rangle \\ &= F_{ji}^T, \qquad \text{for } k+1 \le i \le 2k, \ 1 \le j \le k, \end{aligned} \tag{20}$$

$$\begin{aligned} F_{ij} &= \left\langle \frac{\partial f(\boldsymbol{x},\boldsymbol{\theta})}{\partial w_{i-k}} \frac{\partial f(\boldsymbol{x},\boldsymbol{\theta})}{\partial w_{j-k}} \right\rangle = \langle \phi(\boldsymbol{x},\boldsymbol{J}_{i-k}) \phi(\boldsymbol{x},\boldsymbol{J}_{j-k}) \rangle \\ &= Q_3(\boldsymbol{J}_{i-k},\boldsymbol{J}_{j-k}), \ \text{for } k+1 \le i \le 2k, \ k+1 \le j \le 2k. \end{aligned} \tag{21}$$

Here, we have obtained the analytical form of FIM for MLPs. □

III. Simulation experiments

In this section, we verify the correctness of above theoretical analysis results by taking three experiments. Since calculating the FIM only needs to know the student parameters, the teacher parameters do not play key roles. For convenience and without loss of generality, the teacher model is chosen to be described by MLPs. [11] concluded that it is enough to capture the essence of singular learning dynamics of feedforward neural networks by investigating two hidden units case, thus we choose the teacher model and student model that both have two hidden units, namely the teacher model is:

$$f_0(\boldsymbol{x}) = f(\boldsymbol{x},\boldsymbol{\theta}_0) = v_1 \phi(\boldsymbol{x},\boldsymbol{t}_1) + v_2 \phi(\boldsymbol{x},\boldsymbol{t}_2) + \varepsilon, \tag{22}$$

where $\boldsymbol{\theta_0} = \{\boldsymbol{t}_1, \boldsymbol{t}_2, v_1, v_2\}$ represents all the teacher parameters.

The student model is:

$$f(\boldsymbol{x},\boldsymbol{\theta}) = w_1 \phi(\boldsymbol{x},\boldsymbol{J}_1) + w_2 \phi(\boldsymbol{x},\boldsymbol{J}_2), \tag{23}$$

where $\boldsymbol{\theta} = \{\boldsymbol{J}_1, \boldsymbol{J}_2, w_1, w_2\}$ represents all the student parameters.

As the FIM degenerates on the singularities, thus in order to analyze the case that the learning process has been affected by singularities, we need to choose an index to show whether the FIM degenerates. Here, we choose the inverse of condition value as the index, if the matrix is near singular, the condition value will become very large, i.e. the inverse of condition value will be near zero, otherwise, the inverse of condition value will remain non zero value. For the opposite singularity, we choose the index $h(1,2) = \frac{1}{2}\|\boldsymbol{J}_1 + \boldsymbol{J}_2\|^2$ to directly show whether the student model has been affectd by the opposite singularity. $h(1,2) = 0$ is equivalent to $\boldsymbol{J}_1 = -\boldsymbol{J}_2$, namely the parameters are on the opposite singularity.

Then we take three experiments to verify the correctness of Theorem 1, which include three cases: opposite singularity case, elimination singularity case, fast convergence case that the learning process is not affected by singularities. For a given teacher model, by choosing different initial values of student model, we complete the experiments by using batch mode learning.

The teacher model is chosen as: $\boldsymbol{t}_1 = [-0.20,\ 0.50]^T$, $v_1 = -0.50$, $\boldsymbol{t}_2 = [0.70,\ 0.40]^T$, $v_2 = 0.30$. The additional noise $\varepsilon \sim N(0, 0.05)$. We generate 200 training examples which are subject to standard Gaussian distribution and the learning rate $\eta = 0.005$.

The experiment results are shown as follows:

Case 1 (**Opposite singularity**): the learning process is affected by the opposite singularity.

In this case, the learning dynamics are affected by the opposite singularity and the two hidden units nearly opposite during the training process. The initial student parameters are $\boldsymbol{J}_1^{(0)} = [0.40,\ 0.10]^T$, $w_1^{(0)} = -0.22$, $\boldsymbol{J}_2^{(0)} = [-0.49,\ 0.35]^T$, $w_2^{(0)} = -1.08$. The final student parameters are $\boldsymbol{J}_1 = [0.2975,\ -0.1528]^T$, $w_1 = -0.1247$, $\boldsymbol{J}_2 = [-0.2886,\ 0.1480]^T$, $w_2 = -1.0306$. The results are shown in Fig. 2, which represent the trajectories of training error, $h(1,2)$, and the log scale of the inverse of the condition number, respectively.

As can be seen from Fig. 2(b) and the final values of $\boldsymbol{J}_1$ and $\boldsymbol{J}_2$, $h(1,2)$ decreases to nearly 0 after the training process starts and then this state remains nearly unchanged till the end of training, i.e. the two hidden units nearly opposite. Thus the learning process has been influenced by opposite singularity. In the meanwhile, it can be seen from Fig. 2(c) that the inverse of condition value is smaller than $10E-10$ during the stage

$$\boldsymbol{F}(\boldsymbol{\theta}) = \left\langle \frac{\partial f(\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial f(\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right\rangle$$
$$= \begin{bmatrix} w_1^2\boldsymbol{Q}_1(\boldsymbol{J}_1,\boldsymbol{J}_1) & \cdots & w_1w_k\boldsymbol{Q}_1(\boldsymbol{J}_1,\boldsymbol{J}_k) & w_1\boldsymbol{Q}_2(\boldsymbol{J}_1,\boldsymbol{J}_1) & \cdots & w_1\boldsymbol{Q}_2(\boldsymbol{J}_1,\boldsymbol{J}_k) \\ \vdots & & \vdots & \vdots & & \vdots \\ w_1w_k\boldsymbol{Q}_1(\boldsymbol{J}_k,\boldsymbol{J}_1) & \cdots & w_k^2\boldsymbol{Q}_1(\boldsymbol{J}_k,\boldsymbol{J}_k) & w_k\boldsymbol{Q}_2(\boldsymbol{J}_k,\boldsymbol{J}_1) & \cdots & w_k\boldsymbol{Q}_2(\boldsymbol{J}_k,\boldsymbol{J}_k) \\ w_1\boldsymbol{Q}_2(\boldsymbol{J}_1,\boldsymbol{J}_1)^T & \cdots & w_k\boldsymbol{Q}_2(\boldsymbol{J}_k,\boldsymbol{J}_1)^T & Q_3(\boldsymbol{J}_1,\boldsymbol{J}_1) & \cdots & Q_3(\boldsymbol{J}_1,\boldsymbol{J}_n) \\ \vdots & & \vdots & \vdots & & \vdots \\ w_1\boldsymbol{Q}_2(\boldsymbol{J}_k,\boldsymbol{J}_1)^T & \cdots & w_k\boldsymbol{Q}_2(\boldsymbol{J}_k,\boldsymbol{J}_k)^T & Q_3(\boldsymbol{J}_1,\boldsymbol{J}_n) & \cdots & Q_3(\boldsymbol{J}_n,\boldsymbol{J}_n) \end{bmatrix}. \tag{16}$$

that the two hidden units nearly opposite. This means that the FIM remains nearly singular as the learning process has been affected by the opposite singularity.

Case 2 (**Elimination singularity**): the learning process is affected by the elimination singularity.

For this case, the learning dynamics of MLPs are affected by the elimination singularity and during the learning process we can observe that one output weights crosses 0. The initial student parameters are $\boldsymbol{J}_1^{(0)} = [0.40,\ 0.10]^T$, $w_1^{(0)} = -0.12$, $\boldsymbol{J}_2^{(0)} = [-0.49,\ 0.35]^T$, $w_2^{(0)} = -0.85$. The final student parameters are $\boldsymbol{J}_1 = [0.6481,\ 0.4148]^T$, $w_1 = 0.3381$, $\boldsymbol{J}_2 = [-0.1905,\ 0.5537]^T$, $w_2 = -0.5022$. The simulation results are shown in Fig. 3, which represent the trajectories of training error, output weight $\boldsymbol{w}$ and the inverse of the condition number, respectively.

From Fig. 3(c), we can see that w_2 crosses 0 in the learning process, i.e. the learning dynamics are affected by the elimination singularity and a plateau phenomenon can be obviously observed in Fig. 3(a). During the stage that the learning process has been affected by elimination singularity, as can be seen in Fig. 3(c), FIM becomes nearly singular and finally becomes regular when the parameters escaped the influence of elimination singularity.

Case 3 (**Fast convergence**): the learning process does not suffer from the influence of the singularities

In this case, MLPs are not affected by singularities and the learning dynamics converge to the global minimum fast. The initial student parameters are $\boldsymbol{J}_1^{(0)} = [0.87,\ 0.75]^T$, $w_1^{(0)} = 0.18$, $\boldsymbol{J}_2^{(0)} = [-0.10,\ 0.24]^T$, $w_2^{(0)} = -0.58$. The final student parameters are $\boldsymbol{J}_1 = [0.7661,\ 0.3917]^T$, $w_1 = 0.2917$, $\boldsymbol{J}_2 = [-0.1817,\ 0.4816]^T$, $w_2 = -0.4944$. The simulation results are shown in Fig. 4, which represent the trajectories of training error, output weights $\boldsymbol{w}$, $h(1,2)$ and inverse of the condition number, respectively.

From Fig. 4(a)-(c), we can see that the learning process has not been affected by singularities. Further as shown Fig. 4(d), the condition value of FIM remains nonzero till the end of training process, i.e. the FIM is always nonsingular.

To sum up, the above results of three cases indicate that: when the learning process is affected by the singularities, the FIM becomes singular, and in other cases, the FIM remains regular. This is consistence with the theoretical analysis and verifies the correctness of the analytical form of FIM which we obtained in Theorem 1.

IV. Conclusions

There exist many singularities in the parameter space of multilayer perceptrons, which seriously affect the performance of MLPs. The singularities are the subspaces of parameter space where the Fisher information matrix is singular, thus FIM is an important fundamental to getting the mechanism of the learning dynamics near singularities. In this paper, for the bipolar-activation-function-based MLPs, by adopting the bipolar error function as the activation function, we obtain the analytical form of FIM, which facilitates us to take further analysis of singular learning dynamics. Finally the correctness of the obtained results are verified by taking three experiments in the simulation part.

V. Appendix A

From euqation (2), we have:

$$y - f_0(\boldsymbol{x}) = \varepsilon \sim \mathcal{N}(0, \sigma_0^2), \tag{A-1}$$

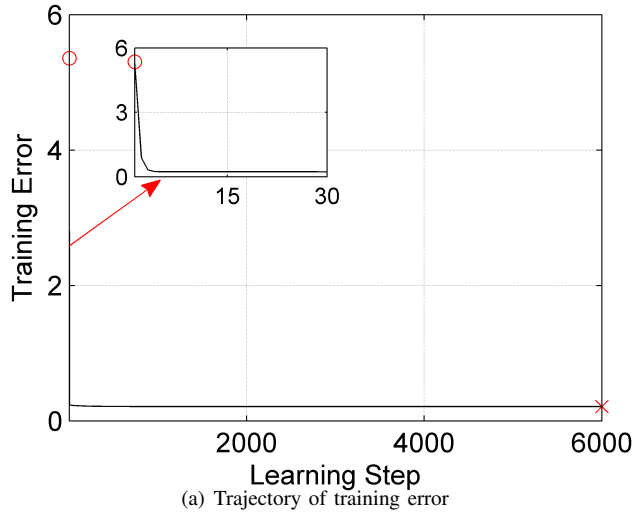
then

$$\frac{1}{\sqrt{2\pi}\sigma_0}\int_{-\infty}^{+\infty} \exp\left(-\frac{(y-f_0(\boldsymbol{x}))^2}{2\sigma_0^2}\right) \mathrm{d}y = \frac{1}{\sqrt{2\pi}\sigma_0}\int_{-\infty}^{+\infty} \exp(-\frac{\varepsilon^2}{2\sigma_0^2})\mathrm{d}\varepsilon = 1. \tag{A-2}$$

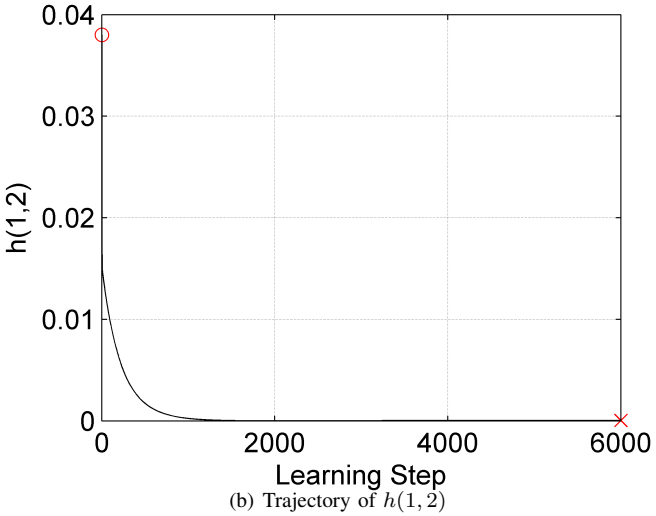
$\boldsymbol{Q}_3(\boldsymbol{J}_i,\boldsymbol{J}_j)$ and $\boldsymbol{Q}_2(\boldsymbol{J}_i,\boldsymbol{J}_j)$ can be rewritten as:

$$\begin{aligned}\boldsymbol{Q}_3(\boldsymbol{J}_i,\boldsymbol{J}_j) &= \left(\sqrt{2\pi}\right)^{-n}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \phi(\boldsymbol{x},\boldsymbol{J}_i)\phi(\boldsymbol{x},\boldsymbol{J}_j) \\ &\quad\times \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{x}\right)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(y-f_0(\boldsymbol{x}))^2}{2}\right)\mathrm{d}y\mathrm{d}\boldsymbol{x} \\ &= \left(\sqrt{2\pi}\right)^{-n}\int_{-\infty}^{+\infty}\phi(\boldsymbol{x},\boldsymbol{J}_i)\phi(\boldsymbol{x},\boldsymbol{J}_j)\exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x}.\end{aligned} \tag{A-3}$$

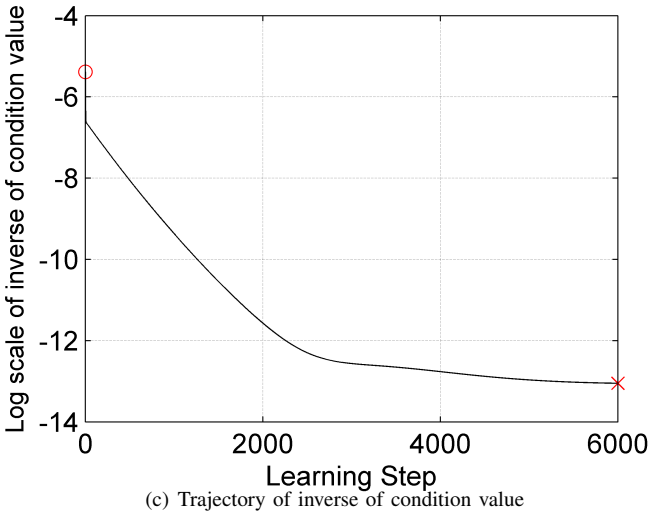
$$\begin{aligned}\boldsymbol{Q}_2(\boldsymbol{J}_i,\boldsymbol{J}_j) &= \left(\sqrt{2\pi}\right)^{-n}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \frac{\partial\phi(\boldsymbol{x},\boldsymbol{J}_i)}{\partial\boldsymbol{J}_i}\phi(\boldsymbol{x},\boldsymbol{J}_j) \\ &\quad\times \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{x}\right)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(y-f_0(\boldsymbol{x}))^2\right)\mathrm{d}y\mathrm{d}\boldsymbol{x} \\ &= \left(\sqrt{2\pi}\right)^{-n}\int_{-\infty}^{+\infty}\frac{\partial\phi(\boldsymbol{x},\boldsymbol{J}_i)}{\partial\boldsymbol{J}_i}\phi(\boldsymbol{x},\boldsymbol{J}_j)\exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x}.\end{aligned} \tag{A-4}$$



(a) Trajectory of training error

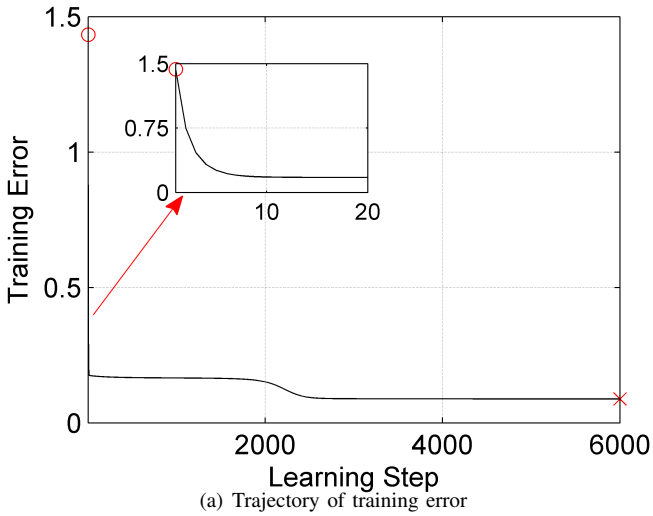


(b) Trajectory of $h(1,2)$

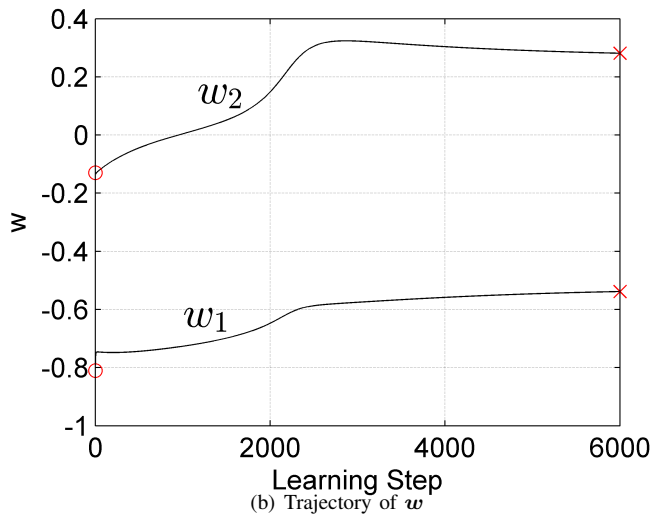


(c) Trajectory of inverse of condition value

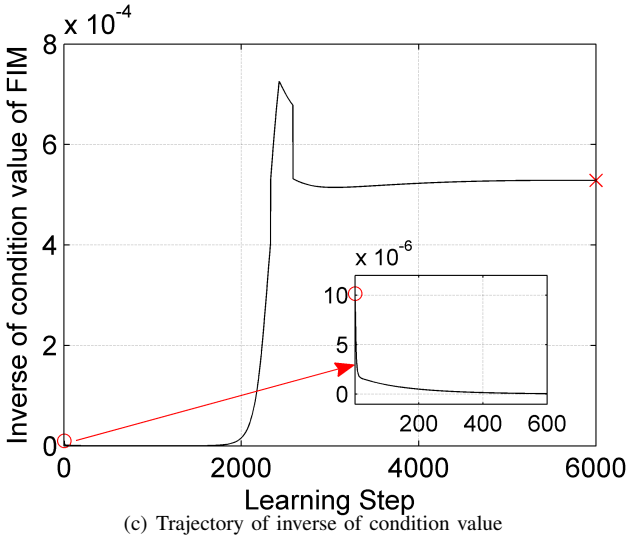
Fig. 2. Case 1 (Opposite singularity) in error function based MLPs



(a) Trajectory of training error

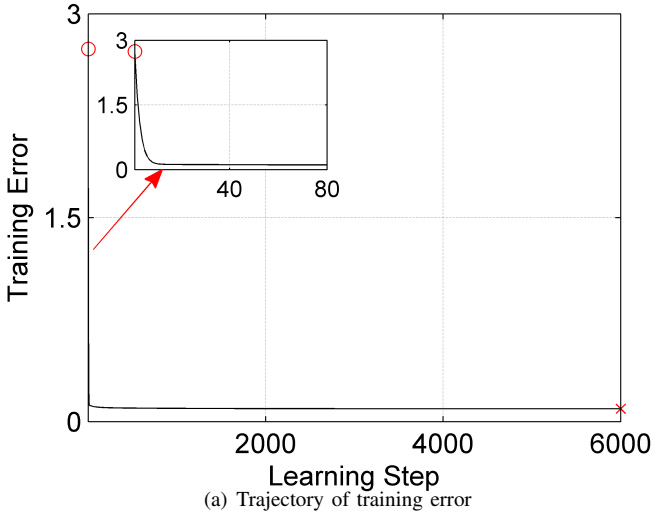


(b) Trajectory of $\boldsymbol{w}$

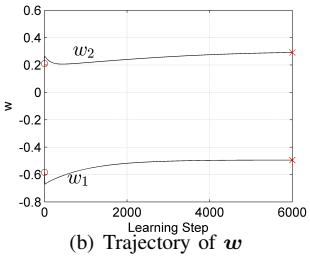


(c) Trajectory of inverse of condition value

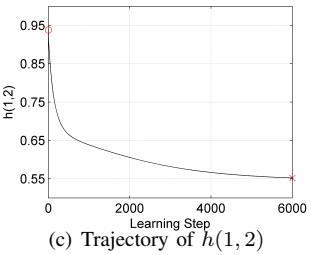
Fig. 3. Case 2 (Elimination singularity) in error function based MLPs



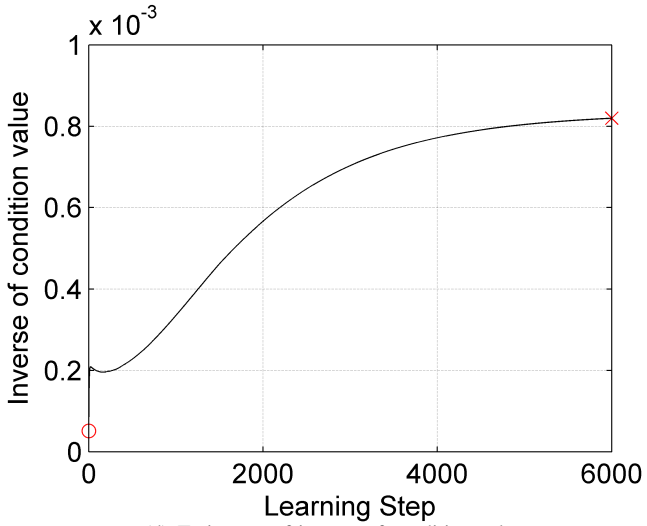
(a) Trajectory of training error



(b) Trajectory of $\boldsymbol{w}$



(c) Trajectory of $h(1,2)$



(d) Trajectory of inverse of condition value

Fig. 4. Case 3 (Fast convergence) in error function based MLPs

Then we have:

$$Q_2(\boldsymbol{J}_i,\boldsymbol{J}_j)=\left(\sqrt{2\pi}\right)^{-n}\int_{-\infty}^{+\infty}\frac{\partial\phi(\boldsymbol{x},\boldsymbol{J}_i)}{\partial\boldsymbol{J}_i}\phi(\boldsymbol{x},\boldsymbol{J}_j)\times\exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x}$$
$$=\sqrt{\frac{2}{\pi}}\left(\sqrt{2\pi}\right)^{-n}\int_{-\infty}^{+\infty}\boldsymbol{x}\exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{J}_i\boldsymbol{J}_i^T\boldsymbol{x}\right)\times\phi(\boldsymbol{x},\boldsymbol{J}_j)\exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x}$$
$$=\left(\sqrt{2\pi}\right)^{-n}\sqrt{\frac{2}{\pi}}\int_{-\infty}^{+\infty}\boldsymbol{x}\phi(\boldsymbol{x},\boldsymbol{J}_j)\times\exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}(\boldsymbol{J}_i)\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x}$$
$$=\sqrt{\frac{2}{\pi}}\frac{\boldsymbol{A}^{-1}(\boldsymbol{J}_i)}{\left(\sqrt{2\pi}\right)^n}\int_{-\infty}^{+\infty}\phi(\boldsymbol{x},\boldsymbol{J}_j)\mathrm{d}\exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}(\boldsymbol{J}_i)\boldsymbol{x}\right)$$
$$=\frac{2}{\pi}\frac{\boldsymbol{A}^{-1}(\boldsymbol{J}_i)}{\left(\sqrt{2\pi}\right)^n}\boldsymbol{J}_j\times\int_{-\infty}^{+\infty}\exp\left(-\frac{1}{2}\boldsymbol{x}^T(\boldsymbol{A}(\boldsymbol{J}_i)+\boldsymbol{J}_j\boldsymbol{J}_j^T)\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x}$$
$$=\frac{2}{\pi}\frac{\boldsymbol{A}^{-1}(\boldsymbol{J}_i)}{\left(\sqrt{2\pi}\right)^n}\boldsymbol{J}_j\int_{-\infty}^{+\infty}\exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{B}(\boldsymbol{J}_i,\boldsymbol{J}_j)\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x}$$
$$=\frac{2}{\pi}\sqrt{\det(\boldsymbol{B}^{-1}(\boldsymbol{J}_i,\boldsymbol{J}_j))}\boldsymbol{A}^{-1}(\boldsymbol{J}_i)\boldsymbol{J}_j, \quad \text{(A-5)}$$

where

$$\boldsymbol{A}(\boldsymbol{J}_i)=\mathbf{I}_n+\boldsymbol{J}_i\boldsymbol{J}_i^T, \quad \text{(A-6)}$$

$$\boldsymbol{B}(\boldsymbol{J}_i,\boldsymbol{J}_j)=\boldsymbol{A}(\boldsymbol{J}_i)+\boldsymbol{J}_j\boldsymbol{J}_j^T=\mathbf{I}_n+\boldsymbol{J}_i\boldsymbol{J}_i^T+\boldsymbol{J}_j\boldsymbol{J}_j^T, \quad \text{(A-7)}$$

$\mathbf{I}_n$ is the compatible identity matrix.

According to Sherman-Morrison formula, we have:

$$\boldsymbol{A}^{-1}(\boldsymbol{J}_i)=\left(\mathbf{I}_n+\boldsymbol{J}_i\boldsymbol{J}_i^T\right)^{-1}=\mathbf{I}_n-\frac{\boldsymbol{J}_i\boldsymbol{J}_i^T}{1+\|\boldsymbol{J}_i\|^2}, \quad \text{(A-8)}$$

$$\mathbf{B}^{-1}(\boldsymbol{J}_i,\boldsymbol{J}_j)=\left(\boldsymbol{A}(\boldsymbol{J}_i)+\boldsymbol{J}_j\boldsymbol{J}_j^T\right)^{-1}=\boldsymbol{A}^{-1}(\boldsymbol{J}_i)-\frac{\boldsymbol{A}^{-1}(\boldsymbol{J}_i)\boldsymbol{J}_j\boldsymbol{J}_j^T\boldsymbol{A}^{-1}(\boldsymbol{J}_i)}{1+\boldsymbol{J}_j^T\boldsymbol{A}^{-1}(\boldsymbol{J}_i)\boldsymbol{J}_j}. \quad \text{(A-9)}$$

By using the matrix determinant lemma, we have:

$$\det(\boldsymbol{A}(\boldsymbol{J}_i))=\det(\boldsymbol{I}+\boldsymbol{J}_i\boldsymbol{J}_i^T)=1+\boldsymbol{J}_i^T\boldsymbol{J}_i, \quad \text{(A-10)}$$

$$\det(\boldsymbol{A}^{-1}(\boldsymbol{J}_i))=\frac{1}{\det(\boldsymbol{A}(\boldsymbol{J}_i))}=\frac{1}{1+\boldsymbol{J}_i^T\boldsymbol{J}_i}, \quad \text{(A-11)}$$

$$\det(\boldsymbol{B}(\boldsymbol{J}_i,\boldsymbol{J}_j))=\det(\boldsymbol{A}(\boldsymbol{J}_i)+\boldsymbol{J}_j\boldsymbol{J}_j^T)=(1+\boldsymbol{J}_j^T\boldsymbol{A}(\boldsymbol{J}_i)\boldsymbol{J}_j)\det(\boldsymbol{A}(\boldsymbol{J}_i)), \quad \text{(A-12)}$$

$$\det(\boldsymbol{B}^{-1}(\boldsymbol{J}_i,\boldsymbol{J}_j))=\frac{1}{\det(\boldsymbol{B}(\boldsymbol{J}_i,\boldsymbol{J}_j))}=\frac{1}{(1+\boldsymbol{J}_j^T\boldsymbol{A}^{-1}(\boldsymbol{J}_i)\boldsymbol{J}_j)\det(\boldsymbol{A}(\boldsymbol{J}_i))}=\frac{1}{(1+\boldsymbol{J}_i^T\boldsymbol{J}_i)(1+\boldsymbol{J}_j^T\boldsymbol{J}_j)-(\boldsymbol{J}_i^T\boldsymbol{J}_j)^2}. \quad \text{(A-13)}$$

Here the analytical form of $\boldsymbol{Q}_2(\boldsymbol{J}_i,\boldsymbol{J}_j)$ has been obtained. According to the Leibniz integral rule, we can get:

$$\boldsymbol{Q}_2(\boldsymbol{J}_i,\boldsymbol{J}_j)=\frac{\partial Q_3(\boldsymbol{J}_i,\boldsymbol{J}_j)}{\partial\boldsymbol{J}_i}. \quad \text{(A-14)}$$

$Q_3(\boldsymbol{J}_i, \boldsymbol{J}_j)$ can be obtained by integrating $\boldsymbol{Q}_2(\boldsymbol{J}_i, \boldsymbol{J}_j)$ respective to $\boldsymbol{J}_i$, then we can get:

$$
\begin{aligned}
Q_3(\boldsymbol{J}_i, \boldsymbol{J}_j) &= \int_{-\infty}^{\boldsymbol{J}_i} \boldsymbol{Q}_2(\boldsymbol{J}_i, \boldsymbol{J}_j)\,\mathrm{d}\boldsymbol{J}_i \\
&= \int_{-\infty}^{\boldsymbol{J}_i} \frac{2}{\pi}\sqrt{\det(\boldsymbol{B}^{-1}(\boldsymbol{J}_i, \boldsymbol{J}_j^T))}\boldsymbol{A}^{-1}(\boldsymbol{J}_i)\boldsymbol{J}_j\,\mathrm{d}\boldsymbol{J}_i \\
&= \frac{2}{\pi}\sqrt{\det(\boldsymbol{B}^{-1}(\boldsymbol{J}_i, \boldsymbol{J}_j^T))}\int_{-\infty}^{\boldsymbol{J}_i} \frac{1}{\sqrt{1+\boldsymbol{J}_i^T\boldsymbol{J}_i}\sqrt{1+\boldsymbol{J}_j^T\boldsymbol{J}_j}} \\
&\qquad \times \frac{1}{\sqrt{1-\frac{(\boldsymbol{J}_i^T\boldsymbol{J}_j)^2}{(1+\boldsymbol{J}_i^T\boldsymbol{J}_i)(1+\boldsymbol{J}_j^T\boldsymbol{J}_j)}}}\boldsymbol{A}^{-1}(\boldsymbol{J}_i)\boldsymbol{J}_j\,\mathrm{d}\boldsymbol{J}_i \\
&= \frac{2}{\pi}\int_{-\infty}^{\boldsymbol{J}_i} \frac{1}{\sqrt{1-\frac{(\boldsymbol{J}_i^T\boldsymbol{J}_j)^2}{(1+\boldsymbol{J}_i^T\boldsymbol{J}_i)(1+\boldsymbol{J}_j^T\boldsymbol{J}_j)}}} \\
&\qquad \times \frac{\boldsymbol{A}^{-1}(\boldsymbol{J}_i)\boldsymbol{J}_j}{\sqrt{1+\boldsymbol{J}_i^T\boldsymbol{J}_i}\sqrt{1+\boldsymbol{J}_j^T\boldsymbol{J}_j}}\,\mathrm{d}\boldsymbol{J}_i \\
&= \frac{2}{\pi}\times \\
&\int_{-\infty}^{\boldsymbol{J}_i} \frac{1}{\sqrt{1-\frac{(\boldsymbol{J}_i^T\boldsymbol{J}_j)^2}{(1+\boldsymbol{J}_i^T\boldsymbol{J}_i)(1+\boldsymbol{J}_j^T\boldsymbol{J}_j)}}}\,\mathrm{d}\frac{\boldsymbol{J}_i^T\boldsymbol{J}_j}{\sqrt{1+\boldsymbol{J}_i^T\boldsymbol{J}_i}\sqrt{1+\boldsymbol{J}_j^T\boldsymbol{J}_j}} \\
&= \frac{2}{\pi}\left(\arcsin\frac{\boldsymbol{J}_i^T\boldsymbol{J}_j}{\sqrt{1+\boldsymbol{J}_i^T\boldsymbol{J}_i}\sqrt{1+\boldsymbol{J}_j^T\boldsymbol{J}_j}} + C_0\right), \qquad \text{(A-15)}
\end{aligned}
$$

where C_0 is a constant.

As $Q_3(\mathbf{0}, \mathbf{0}) = 0$, then we have $C_0 = 0$. Finally we get

$$
Q_3(\boldsymbol{J}_i, \boldsymbol{J}_j) = \frac{2}{\pi}\arcsin\frac{\boldsymbol{J}_i^T\boldsymbol{J}_j}{\sqrt{1+\boldsymbol{J}_i^T\boldsymbol{J}_i}\sqrt{1+\boldsymbol{J}_j^T\boldsymbol{J}_j}}. \qquad \text{(A-16)}
$$

For $\boldsymbol{Q}_1(\boldsymbol{J}_i, \boldsymbol{J}_j)$, we have:

$$
\begin{aligned}
&\boldsymbol{Q}_1(\boldsymbol{J}_i, \boldsymbol{J}_j) = \\
&\left(\sqrt{2\pi}\right)^{-n}\int_{-\infty}^{+\infty}\frac{\partial\phi(\boldsymbol{x},\boldsymbol{\theta})}{\partial\boldsymbol{J}_i}\frac{\partial\phi(\boldsymbol{x},\boldsymbol{\theta})}{\partial\boldsymbol{J}_j^T}\exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x} \\
&= \frac{2}{\pi}\left(\sqrt{2\pi}\right)^{-n}\int_{-\infty}^{+\infty}\boldsymbol{x}\boldsymbol{x}^T\exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{J}_i\boldsymbol{J}_i^T\boldsymbol{x}\right) \\
&\qquad \times\exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{J}_j\boldsymbol{J}_j^T\boldsymbol{x}\right)\exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x} \\
&= \frac{2}{\pi}\left(\sqrt{2\pi}\right)^{-n}\int_{-\infty}^{+\infty}\boldsymbol{x}\boldsymbol{x}^T\exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{B}(\boldsymbol{J}_i,\boldsymbol{J}_j)\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x} \\
&= \frac{2}{\pi}\sqrt{\det(\boldsymbol{B}^{-1}(\boldsymbol{J}_i,\boldsymbol{J}_j))}\frac{\left(\sqrt{2\pi}\right)^{-n}}{\sqrt{\det(\boldsymbol{B}^{-1}(\boldsymbol{J}_i,\boldsymbol{J}_j))}} \\
&\quad \times\int_{-\infty}^{+\infty}\boldsymbol{x}\exp\left(-\frac{1}{2}\left(\boldsymbol{x}^T\left(\boldsymbol{B}^{-1}(\boldsymbol{J_i},\boldsymbol{J_j})\right)^{-1}\boldsymbol{x}\right)\right)\boldsymbol{x}^T\mathrm{d}\boldsymbol{x} \\
&= \frac{2}{\pi}\sqrt{\det(\boldsymbol{B}^{-1}(\boldsymbol{J}_i,\boldsymbol{J}_j))}\boldsymbol{B}^{-1}(\boldsymbol{J_i},\boldsymbol{J_j}). \qquad \text{(A-17)}
\end{aligned}
$$

□

REFERENCES

[1] C. Mu, Z. Ni, C. Sun, H. He. "Data-driven tracking control with adaptive dynamic programming for a class of continuous-time nonlinear systems," *IEEE Transactions on Cybernetics*, vol. 47, no. 6, pp. 1460–1470, 2017.

[2] L. Xie, D. Tao, H. Wei. "Early Expression detection via online multi-instance learning with nonlinear extension," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 5, pp. 1486–1496, 2019.

[3] J. Zhang, K A. Ehinger, H. Wei, et al. "A novel graph-based optimization framework for salient object detection," *Pattern Recognition*, vol.64, pp. 39–50, 2017.

[4] M. Biehl, H. Schwarze. "Learning by on-line gradient descent. *Journal of Physics A: Mathematical and General*, vol. 28, No. 3, pp. 643–656, 1995.

[5] K. Fukumizu, S. Amari. "Local minima and plateaus in hierarchical structure of multilayer perceptrons," *Neural Networks*, vol. 13, No. 3, pp. 317–327, 2000.

[6] S. Amari, T. Ozeki. "Differential and algebraic geometry of multilayer perceptrons," *IEICE Trans. Fundamentals of Electronics Communications and Computer Sciences*, vol. E84-A, pp. 31–38, 2001.

[7] S. Amari, H. Nagaoka," *Information geometry*. New York: AMS and Oxford University Press, 2000.

[8] S. Watanabe. "Almost All Learning Machines are Singular," *in Proceedings of IEEE Symposium on Foundations of Computational Intelligence*, pp. 383–388, 2007.

[9] H. Wei, S. Amari. "Eigenvalue analysis on singularity in RBF networks," *in Proceedings of International Joint Conference on Neural Networks*, pp. 690–695, 2007.

[10] S. Amari, H. Park, T. Ozeki. "Singularities affect dynamics of learning in neuromanifolds. *Neural Computation*. 18(5): 1007–1065, 2006.

[11] H. Wei, J. Zhang, F. Cousseau, T. Ozeki, S. Amari. "Dynamics of learning near singularities in layered networks," *Neural Computation*, vol. 20, No. 3, pp. 813–843, 2008.

[12] W. Guo, H. Wei, J. Zhao, K. Zhang. "Theoretical and numerical analysis of learning dynamics near singularity in multilayer perceptrons," *Neurocomputing*, vol. 151, pp. 390–400, 2015.

[13] H. Wei, S. Amari. "Dynamics of learning near singularities in radial basis function networks," *Neural Networks*, vol. 21, No.7, pp. 989–1005, 2008.

[14] H. Park, T. Ozeki. "Singularity and Slow Convergence of the EM algorithm for Gaussian Mixtures," *Neural Process Letters*, vol. 29, No. 1, pp. 45–59, 2009.

[15] F. Cousseau, T. Ozeki, S. Amari. "Dynamics of learning in multilayer perceptrons near singularities," *IEEE Transaction on Neural Networks*, vol. 19, No. 8, pp. 1313–1328, 2008.

[16] W. Guo, J. Zhao, J. Zhang, H. Wei, A. Song, K. Zhang. "Stability analysis of opposite singularity in multilayer perceptrons," *Neurocomputing*, vol. 282, pp. 192–201, 2018.

[17] H. Park, M. Inoue, M. Okada. “Online learning dynamics of multilayer perceptrons with unidentifiable parameters,” *Journal of Physics A: Mathematical and General*, vol. 36, No. 47, pp. 11753–11764, 2003.

[18] S. Amari. “Neural learning in structured parameter spaces-natural Riemannian gradient, *in Proceedings of 10th Advances in Neural Information Processing Systems*, pp. 127–133, 1997.

[19] S. Amari. “Natural gradient works efficiently in learning,” *Neural computation*, vol. 10, No. 2, pp. 251–276, 1998.

[20] H. Park, S. Amari, K. Fukumizu. “Adaptive natural gradient learning algorithms for various stochastic models,” *Neural Networks*, vol. 13, No. 7, pp. 755–764, 2000.

[21] J. Zhao, H. Wei, C. Zhang, et al. “Natural gradient learning algorithms for RBF networks,” *Neural Computation*, vol. 27, No. 2, pp. 481–505, 2015.

[22] R. Grosse, R. Salakhudinov. “Scaling up natural gradient by sparsely factorizing the inverse fisher matrix,” *in Proceedings of the 32nd International Conference on Machine Learning (ICML2015)*, pp. 2304–2313, 2015.

[23] Park H, Lee K. “Adaptive natural gradient method for learning neural networks with large data set in mini-batch mode,” *in Proceeding of International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pp. 306–310, 2019.