IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 5 November 2022, accepted 26 November 2022, date of publication 7 December 2022,
date of current version 13 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3227427

RESEARCH ARTICLE

Fisher Information Matrix and its Application of Bipolar Activation Function Based Multilayer Perceptrons With General Gaussian Input

WEILI GUO[1,2,3], GUANGYU LI[1,2,3], AND JIANFENG LU[1], (Member, IEEE)

[1] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
[2] Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, Nanjing University of Science and Technology, Nanjing 210094, China
[3] Jiangsu Key Laboratory of Image and Video Understanding for Social Security, Nanjing University of Science and Technology, Nanjing 210094, China

Corresponding author: Guangyu Li (guangyu.li2017@njust.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61906092, Grant 61802059, Grant 62006119, and Grant 61876085; in part by the Natural Science Foundation of Jiangsu Province of China under Grant BK20190441, Grant BK20180365, and Grant BK20190444; in part by the China Postdoctoral Science Foundation under Grant 2021M691596.

ABSTRACT For the widely used multilayer perceptrons (MLPs), there exist singularities in the parameter space where Fisher information matrix (FIM) degenerates on these subspaces. The singularities seriously influence the learning dynamics of MLPs which have attracted many researchers' attentions. As FIM plays key role in investigating the singular learning dynamics of MLPs, it is very important to obtain the analytical form of FIM. In this paper, for the bipolar activation function based MLPs with general Gaussian input, by choosing bipolar error function as the activation function, the analytical form of FIM are obtained. Then the validity of obtained results are verified by taking two experiments.

INDEX TERMS Fisher information matrix, multilayer perceptrons, singularity, bipolar error function, general Gaussian input.

I. INTRODUCTION

As one of the most important subject in computer science, artificial intelligence has been developed fast in the last years and has been successfully applied in various areas and applications [1], [2], such as pattern recognition, computer vision, intelligence control etc [3], [4], [5]. For artificial intelligence, artificial neural networks play key roles in achieving such outstanding performance [6], [7]. Multilayer perceptrons (MLPs), which are typical feedforward neural networks, also have been widely applied in artificial intelligence [8], [9]. The main advantages of multilayer perceptrons are that they are easy to handle and can approximate any continuous function arbitrary well.

However, different with the regular learning machines, when researchers used MLPs to different applications, they found that there were some strange behaviours in the learning process of MLPs [10]. For example, there are many local minima, the learning process may become very slow and the so-called plateau phenomenon can often be observed (an example is shown in Fig. 1) [11]. In view of the wide applications of MLPs, the reasons why the training processes often suffer from such difficulties have attracted many researchers' attentions. Research results indicate that these singular behaviours are because of the network structure of feedforward neural networks which have hidden layers. Due to the existence of hidden layers, there exist subspaces in the parameter space of feedforward neural networks where the Fisher information matrix (FIM) is singular on such subspaces [12], [13]. These subspaces mainly cause the above singular learning behaviours of MLPs, thus we call these subspaces as singularities.

As the FIM degenerates on singularities, the subspaces become Riemann manifolds, not Euclidean spaces in the case of regular learning machines, which leads to three problems [11], [14]: 1) invalidation of the classic paradigm of Cramer-Rao theorem; 2) failure to determine approximate network structure. For example, for the commonly used

The associate editor coordinating the review of this manuscript and approving it for publication was Gerardo Flores.

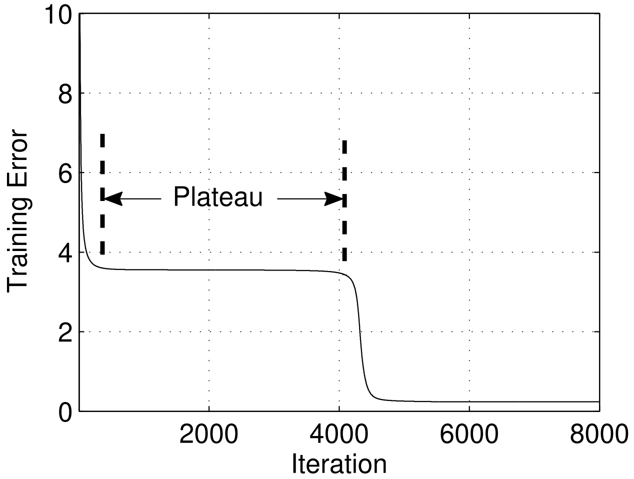


FIGURE 1. Plateau phenomenon occurred in the learning process of MLPs.

model selection criteria, such as Akaike information criterion (AIC), Bayes information criterion (BIC) and minimum description length (MDL), researchers find that these criteria often fail to determine approximate network structure; 3) non Fisher-efficiency of standard gradient descent method. Instead of gradient descent direction, the Riemann gradient (natural gradient) descent direction becomes the steepest descent direction [15], then using standard gradient descent method to train neural networks will face many difficulties on the singularities. Therefore, it is very worthy to investigate the learning dynamics near singularities in MLPs.

Given that FIM plays fundamental and vital role in investigating the singular learning dynamics of MLPs, obtaining the analytical form of FIM has two important significances: 1) make us convenient to detailed analyze the mechanism of singular learning dynamics; 2) make it easier to design better learning algorithms to overcome the serious influence of singularities. Thus the main contribution of this paper is to obtain the analytical form of FIM for the bipolar-error-function-based MLPs with general Gaussian input. Further we also show the potential of analytical form to design better algorithms.

The rest of this paper is organized as follows. A brief review of related work is presented in section 2. In section 3, the analytical form of FIM is obtained. In section 4, we verify the validity of the obtained results through simulation studies. Section 5 states conclusions and discussions.

II. RELATED WORK

In this section, we provide a brief overview of previous work on the mechanism of singular learning dynamics.

By investigating the geometric structure of MLPs, [16] proved that the global minimum of the smaller model could be a local minimum or a saddle point of the larger model and illustrated various singularities in detail. For layered networks, by taking general mathematical analysis, [17] obtained universal learning trajectories near the overlap singularity. Further researchers aimed to take more detailed theoretical analysis on the learning dynamics near singularities. However, the widely used activation functions, such as log-sigmoid function $\frac{1}{1+e^{-\lambda x}}$ and hyperbolic tangent function $\tanh(x)$, can not be integrated, which limits researchers to take quantitative analysis of learning dynamics. In order to overcome this problem, the error functions $\phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left(-\frac{1}{2}t^2\right)\mathrm{d}t$ and $\phi(x) = \sqrt{\frac{2}{\pi}}\int_{0}^{x}\exp\left(-\frac{1}{2}t^2\right)\mathrm{d}t$, were chosen as the activation function of MLPs in unipolar and bipolar case, respectively [11], [18]. Then different cases of MLPs with different type of activation functions, including toy model case [19], regular case [20], [21], and unrealizable case [22], have been investigated and diverse results have also been obtained. [23] obtained the analytical form of FIM in RBF networks and investigated to what extent RBF networks would be influenced by singularities.

Since the Riemann gradient (natural gradient) descent direction becomes the steepest descent direction on the singularities, the natural gradient method was proposed to overcome the serious influence of singualarities [24]. As it is very hard to obtain the explicit form of FIM and its inverse, researchers proposed adaptive natural gradient algorithms where the inverse FIM is calculated by directly using approximation formula [25], [26], [27] and applied natural gradient method in big data fields and deep neural networks [28], [29], [30].

Due to the non-integrated property of hyperbolic tangent function, we cannot obtain the analytical form of FIM. In this paper, we choose the bipolar error function $\phi(x) = \sqrt{\frac{2}{\pi}}\int_{0}^{x}\exp\left(-\frac{1}{2}t^2\right)\mathrm{d}t$ as the the activation function of MLPs, and obtain the analytical form of FIM.

III. ANALYTICAL FORM OF FISHER INFORMATION MATRIX

In this section, the learning paradigm of MLPs is introduced at first and then the analytical form of FIM is obtained.

The bipolar-activation-function based multilayer perceptrons with one hidden layer are defined as follows:

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

where $\boldsymbol{x}$ is the input, k is the hidden node number, $\boldsymbol{J}_i$ and w_i are the weight from input layer to hidden node i and weight from hidden node i to output layer, respectively. $\phi(\cdot)$ is a bipolar activation function. Then $\boldsymbol{\theta} = \{\boldsymbol{J}_1,\cdots,\boldsymbol{J}_k,w_1,\cdots,w_k\}$ represents all the parameters of the model. In order to obtain the analytical form of FIM and overcome the non-integrated property of hyperbolic tangent function, in this paper, we choose the bipolar error function as the activation function, namely $\phi(\boldsymbol{x},\boldsymbol{J}_i) = \sqrt{\frac{2}{\pi}}\int_{0}^{\boldsymbol{J}_i^T\boldsymbol{x}}\exp\left(-\frac{1}{2}t^2\right)\mathrm{d}t$.

For the regression mission, an unknown teacher function is needed to be approximated:

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

which generates a number of observed data $(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_t, y_t)$. The additive noise ε usually subjects to a Gaussian distribution with mean 0 and variance σ_0^2.

Generally the input $\boldsymbol{x}$ is assumed to be subject to Gaussian distribution, in this paper, we investigate the general Gaussian input case, i.e. probability density function of $\boldsymbol{x}$ is:

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

where n is the input dimension, $\boldsymbol{\mu}$ is the expectation value and $\boldsymbol{\Sigma}$ is the covariance matrix.

We choose the square loss function to measure the error:

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

and use the gradient descent method to minimize the 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 is the learning rate.

The FIM is defined as follows [11]:

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

where $\langle\cdot\rangle$ denotes the expectation with respect to the teacher distribution. The teacher distribution is given by:

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

Then we introduce the types of singularities. As shown in [11], besides of the overlap singularity and elimination singularity in the parameter space of unipolar-activation-function-based MLPs, there also exists opposite singularity for the bipolar-activation-function-based MLPs (1), thus there are total three types of singularities:

(1) Opposite singularity:

$$\mathcal{R}_1 = \{\boldsymbol{\theta} | \boldsymbol{J}_i = -\boldsymbol{J}_j\}, \tag{8}$$

(2) Overlap singularity:

$$\mathcal{R}_2 = \{\boldsymbol{\theta} | \boldsymbol{J}_i = \boldsymbol{J}_j\}, \tag{9}$$

(3) Elimination singularity:

$$\mathcal{R}_3 = \{\boldsymbol{\theta} | w_i = 0\}. \tag{10}$$

Now we aim to obtain the explicit expression of FIM. For the Gaussian input case, the covariation matrix $\boldsymbol{\Sigma}$ plays a center role and the value of $\boldsymbol{\mu}$ does not essentially influence on the analytical process, without loss of generality, $\boldsymbol{\mu}$ is adopted as $\boldsymbol{0}$ in this paper.

Before we give the analytical form of FIM, we firstly obtain the explicit expressions of $\left\langle \frac{\partial \phi(\boldsymbol{x}, \boldsymbol{J}_i)}{\partial \boldsymbol{J}_i} \frac{\partial \phi(\boldsymbol{x}, \boldsymbol{J}_j)}{\partial \boldsymbol{J}_j^T} \right\rangle$, $\left\langle \frac{\partial \phi(\boldsymbol{x}, \boldsymbol{J}_i)}{\partial \boldsymbol{J}_i} \phi(\boldsymbol{x}, \boldsymbol{J}_j) \right\rangle$, and $\langle \phi(\boldsymbol{x}, \boldsymbol{J}_i)\phi(\boldsymbol{x}, \boldsymbol{J}_j)\rangle$, which play key role in obtaining the analytical form of FIM. For simplicity, we note:

$$\boldsymbol{Q}_1(\boldsymbol{J}_i, \boldsymbol{J}_j) = \left\langle \frac{\partial \phi(\boldsymbol{x}, \boldsymbol{J}_i)}{\partial \boldsymbol{J}_i} \frac{\partial \phi(\boldsymbol{x}, \boldsymbol{J}_j)}{\partial \boldsymbol{J}_j^T} \right\rangle. \tag{11}$$

$$\boldsymbol{Q}_2(\boldsymbol{J}_i, \boldsymbol{J}_j) = \left\langle \frac{\partial \phi(\boldsymbol{x}, \boldsymbol{J}_i)}{\partial \boldsymbol{J}_i} \phi(\boldsymbol{x}, \boldsymbol{J}_j) \right\rangle. \tag{12}$$

$$Q_3(\boldsymbol{J}_i, \boldsymbol{J}_j) = \langle \phi(\boldsymbol{x}, \boldsymbol{J}_i)\phi(\boldsymbol{x}, \boldsymbol{J}_j)\rangle. \tag{13}$$

Then in Lemma 1, we give the explicit expressions of Eqs. (11)-(13).

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

$$\boldsymbol{Q}_1(\boldsymbol{J}_i, \boldsymbol{J}_j) = \frac{2}{\pi}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}|\boldsymbol{A}(\boldsymbol{J}_i, \boldsymbol{J}_j)|^{\frac{1}{2}}\boldsymbol{A}(\boldsymbol{J}_i, \boldsymbol{J}_j), \tag{14}$$

$$\boldsymbol{Q}_2(\boldsymbol{J}_i, \boldsymbol{J}_j) = \frac{2}{\pi}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}|\boldsymbol{A}(\boldsymbol{J}_i, \boldsymbol{J}_j)|^{\frac{1}{2}}\boldsymbol{B}(\boldsymbol{J}_i)^{-1}\boldsymbol{J}_j, \tag{15}$$

$$Q_3(\boldsymbol{J}_i, \boldsymbol{J}_j) = \frac{2}{\pi}\arcsin\frac{\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_j}{\sqrt{1+\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_i}\sqrt{1+\boldsymbol{J}_j^T\boldsymbol{\Sigma}\boldsymbol{J}_j}}, \tag{16}$$

where:

$$\begin{aligned}\boldsymbol{A}(\boldsymbol{J}_i, \boldsymbol{J}_j) &= (\boldsymbol{\Sigma}^{-1} + \boldsymbol{J}_i\boldsymbol{J}_i^T + \boldsymbol{J}_j\boldsymbol{J}_j^T)^{-1} = \boldsymbol{\Sigma} \\ &\quad -\boldsymbol{\Sigma}\left(\frac{(1+\boldsymbol{J}_j^T\boldsymbol{\Sigma}\boldsymbol{J}_j)\boldsymbol{J}_i\boldsymbol{J}_i^T + (1+\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_i)\boldsymbol{J}_j\boldsymbol{J}_j^T}{(1+\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_i)(1+\boldsymbol{J}_j^T\boldsymbol{\Sigma}\boldsymbol{J}_j) - (\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_j)^2}\right. \\ &\quad \left. - \frac{\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_j(\boldsymbol{J}_i\boldsymbol{J}_j^T + \boldsymbol{J}_j\boldsymbol{J}_i^T)}{(1+\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_i)(1+\boldsymbol{J}_j^T\boldsymbol{\Sigma}\boldsymbol{J}_j) - (\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_j)^2}\right)\boldsymbol{\Sigma},\end{aligned} \tag{17}$$

$$|\boldsymbol{A}(\boldsymbol{J}_i, \boldsymbol{J}_j| = \frac{|\boldsymbol{\Sigma}|}{(1+\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_i)(1+\boldsymbol{J}_j^T\boldsymbol{\Sigma}\boldsymbol{J}_j) - (\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_j)^2}, \tag{18}$$

$$\boldsymbol{B}(\boldsymbol{J}_i) = \boldsymbol{\Sigma}^{-1} + \boldsymbol{J}_i\boldsymbol{J}_i^T, \tag{19}$$

$$\boldsymbol{B}(\boldsymbol{J}_i)^{-1} = \left(\boldsymbol{\Sigma}^{-1} + \boldsymbol{J}_i\boldsymbol{J}_i^T\right)^{-1} = \boldsymbol{\Sigma} - \frac{\boldsymbol{\Sigma}\boldsymbol{J}_i\boldsymbol{J}_i^T\boldsymbol{\Sigma}}{1+\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_i}. \tag{20}$$

Proof: We present the calculation processing in Appendix. □

Now we can give the analytical form of FIM in Theorem 1.

Theorem 1: The analytical form of FIM $\boldsymbol{F}(\boldsymbol{\theta})$ is given by:

$$\boldsymbol{F}(\boldsymbol{\theta}) = \left\langle \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right\rangle = \begin{bmatrix} \boldsymbol{F}_1(\boldsymbol{\theta}) & \boldsymbol{F}_2(\boldsymbol{\theta}) \\ \boldsymbol{F}_3(\boldsymbol{\theta}) & \boldsymbol{F}_4(\boldsymbol{\theta}) \end{bmatrix},$$

where:

$$\boldsymbol{F}_1(\boldsymbol{\theta}) = \begin{bmatrix} w_1^2\boldsymbol{Q}_1(\boldsymbol{J}_1, \boldsymbol{J}_1) & \cdots & w_1w_k\boldsymbol{Q}_1(\boldsymbol{J}_1, \boldsymbol{J}_k) \\ w_1w_2\boldsymbol{Q}_1(\boldsymbol{J}_1, \boldsymbol{J}_2)^T & \cdots & w_2w_k\boldsymbol{Q}_1(\boldsymbol{J}_2, \boldsymbol{J}_k) \\ \vdots & \vdots & \vdots \\ w_1w_k\boldsymbol{Q}_1(\boldsymbol{J}_1, \boldsymbol{J}_k)^T & \cdots & w_k^2\boldsymbol{Q}_1(\boldsymbol{J}_k, \boldsymbol{J}_k) \end{bmatrix}, \tag{21}$$

$$\boldsymbol{F}_2(\boldsymbol{\theta}) = \begin{bmatrix} w_1\boldsymbol{Q}_2(\boldsymbol{J}_1, \boldsymbol{J}_1) & \cdots & w_1\boldsymbol{Q}_2(\boldsymbol{J}_1, \boldsymbol{J}_k) \\ w_2\boldsymbol{Q}_2(\boldsymbol{J}_1, \boldsymbol{J}_2) & \cdots & w_2\boldsymbol{Q}_2(\boldsymbol{J}_2, \boldsymbol{J}_k) \\ \vdots & \vdots & \vdots \\ w_k\boldsymbol{Q}_2(\boldsymbol{J}_1, \boldsymbol{J}_k) & \cdots & w_k\boldsymbol{Q}_2(\boldsymbol{J}_k, \boldsymbol{J}_k) \end{bmatrix}, \tag{22}$$

$$\boldsymbol{F}_3(\boldsymbol{\theta}) = \boldsymbol{F}_2(\boldsymbol{\theta})^T, \tag{23}$$

$$\boldsymbol{F}_4(\boldsymbol{\theta}) = \begin{bmatrix} Q_3(\boldsymbol{J}_1, \boldsymbol{J}_1) & Q_3(\boldsymbol{J}_1, \boldsymbol{J}_2) & \cdots & Q_3(\boldsymbol{J}_1, \boldsymbol{J}_k) \\ Q_3(\boldsymbol{J}_1, \boldsymbol{J}_2) & Q_3(\boldsymbol{J}_2, \boldsymbol{J}_2) & \cdots & Q_3(\boldsymbol{J}_2, \boldsymbol{J}_k) \\ \vdots & \vdots & & \vdots \\ Q_3(\boldsymbol{J}_1, \boldsymbol{J}_k) & Q_3(\boldsymbol{J}_2, \boldsymbol{J}_k) & \cdots & Q_3(\boldsymbol{J}_k, \boldsymbol{J}_k) \end{bmatrix}. \tag{24}$$

Proof: Firstly we define:

$$\boldsymbol{F}_1(\boldsymbol{\theta}) = \begin{bmatrix} \left\langle \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{J}_1} \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{J}_1^T} \right\rangle & \cdots & \left\langle \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{J}_1} \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{J}_k^T} \right\rangle \\ \left\langle \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{J}_2} \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{J}_1^T} \right\rangle & \cdots & \left\langle \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{J}_2} \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{J}_k^T} \right\rangle \\ \vdots & \vdots & \vdots \\ \left\langle \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{J}_k} \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{J}_1^T} \right\rangle & \cdots & \left\langle \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{J}_k} \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{J}_k^T} \right\rangle \end{bmatrix}, \tag{25}$$

$$\boldsymbol{F}_2(\boldsymbol{\theta}) = \begin{bmatrix} \left\langle \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{J}_1} \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial w_1} \right\rangle & \cdots & \left\langle \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{J}_1} \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial w_k} \right\rangle \\ \left\langle \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{J}_2} \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial w_1} \right\rangle & \cdots & \left\langle \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{J}_2} \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial w_k} \right\rangle \\ \vdots & \vdots & \vdots \\ \left\langle \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{J}_k} \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial w_1} \right\rangle & \cdots & \left\langle \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{J}_k} \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial w_k} \right\rangle \end{bmatrix}, \tag{26}$$

$$\boldsymbol{F}_3(\boldsymbol{\theta}) = \begin{bmatrix} \left\langle \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial w_1} \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{J}_1^T} \right\rangle & \cdots & \left\langle \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial w_1} \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{J}_k^T} \right\rangle \\ \left\langle \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial w_2} \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{J}_1^T} \right\rangle & \cdots & \left\langle \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial w_2} \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{J}_k^T} \right\rangle \\ \vdots & \vdots & \vdots \\ \left\langle \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial w_k} \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{J}_1^T} \right\rangle & \cdots & \left\langle \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial w_k} \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{J}_k^T} \right\rangle \end{bmatrix}, \tag{27}$$

$$\boldsymbol{F}_4(\boldsymbol{\theta}) = \begin{bmatrix} \left\langle \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial w_1} \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial w_1} \right\rangle & \cdots & \left\langle \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial w_1} \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial w_k} \right\rangle \\ \left\langle \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial w_2} \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial w_1} \right\rangle & \cdots & \left\langle \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial w_2} \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial w_k} \right\rangle \\ \vdots & \vdots & \vdots \\ \left\langle \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial w_k} \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial w_1} \right\rangle & \cdots & \left\langle \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial w_k} \frac{\partial f(\boldsymbol{x}, \boldsymbol{\theta})}{\partial w_k} \right\rangle \end{bmatrix}, \tag{28}$$

then from Eq. (1) and Eq. (6), we have

$$\boldsymbol{F}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{F}_1(\boldsymbol{\theta}) & \boldsymbol{F}_2(\boldsymbol{\theta}) \\ \boldsymbol{F}_3(\boldsymbol{\theta}) & \boldsymbol{F}_4(\boldsymbol{\theta}) \end{bmatrix}. \tag{29}$$

For Eqs. (25)-(28), by using the results in Lemma 1, we have:

$$\boldsymbol{F}_1(\boldsymbol{\theta}) = \begin{bmatrix} w_1^2 \boldsymbol{Q}_1(\boldsymbol{J}_1, \boldsymbol{J}_1) & \cdots & w_1 w_k \boldsymbol{Q}_1(\boldsymbol{J}_1, \boldsymbol{J}_k) \\ w_1 w_2 \boldsymbol{Q}_1(\boldsymbol{J}_1, \boldsymbol{J}_2)^T & \cdots & w_2 w_k \boldsymbol{Q}_1(\boldsymbol{J}_2, \boldsymbol{J}_k) \\ \vdots & \vdots & \vdots \\ w_1 w_k \boldsymbol{Q}_1(\boldsymbol{J}_1, \boldsymbol{J}_k)^T & \cdots & w_k^2 \boldsymbol{Q}_1(\boldsymbol{J}_k, \boldsymbol{J}_k) \end{bmatrix}, \tag{30}$$

$$\boldsymbol{F}_2(\boldsymbol{\theta}) = \begin{bmatrix} w_1 \boldsymbol{Q}_2(\boldsymbol{J}_1, \boldsymbol{J}_1) & \cdots & w_1 \boldsymbol{Q}_2(\boldsymbol{J}_1, \boldsymbol{J}_k) \\ w_2 \boldsymbol{Q}_2(\boldsymbol{J}_1, \boldsymbol{J}_2) & \cdots & w_2 \boldsymbol{Q}_2(\boldsymbol{J}_2, \boldsymbol{J}_k) \\ \vdots & \vdots & \vdots \\ w_k \boldsymbol{Q}_2(\boldsymbol{J}_1, \boldsymbol{J}_k) & \cdots & w_k \boldsymbol{Q}_2(\boldsymbol{J}_k, \boldsymbol{J}_k) \end{bmatrix}, \tag{31}$$

$$\boldsymbol{F}_3(\boldsymbol{\theta}) = \boldsymbol{F}_2(\boldsymbol{\theta})^T, \tag{32}$$

$$\boldsymbol{F}_4(\boldsymbol{\theta}) = \begin{bmatrix} Q_3(\boldsymbol{J}_1, \boldsymbol{J}_1) & \cdots & Q_3(\boldsymbol{J}_1, \boldsymbol{J}_k) \\ Q_3(\boldsymbol{J}_1, \boldsymbol{J}_2) & \cdots & Q_3(\boldsymbol{J}_2, \boldsymbol{J}_k) \\ \vdots & \vdots & \vdots \\ Q_3(\boldsymbol{J}_1, \boldsymbol{J}_k) & \cdots & Q_3(\boldsymbol{J}_k, \boldsymbol{J}_k) \end{bmatrix}. \tag{33}$$

Till now, the analytical form of FIM has been obtained. □

IV. SIMULATION EXPERIMENTS

In this section, we take three experiments to illustrate the validity and importance of the obtained results. From Eq. (21), we can see that we only need to know the student parameters to obtain the FIM during training process. Thus the type of teacher model does not play a significant role. For convenience and without loss of generality, we investigate the case that the teacher model also has the form of MLPs, i.e. Eq. (2) can be rewritten as:

$$y = f_0(\boldsymbol{x}) = f(\boldsymbol{x}, \boldsymbol{\theta}_0) + \varepsilon = \sum_{i=1}^{M} v_i \phi(\boldsymbol{x}, \boldsymbol{t}_i) + \varepsilon, \tag{34}$$

where M is the hidden unit number.
$\boldsymbol{\theta}_0 = \{\boldsymbol{t}_1, \cdots, \boldsymbol{t}_M, v_1, \cdots, v_M\}$ represents all the teacher parameters. As can be noticed, this assumption is based on the universal approximation ability and is reasonable.

Now we introduce three indexes which are very important to show the experiment results:

1) inverse condition value of FIM

This index is used to judge whether the FIM is singular. When the matrix is nearly singular, the condition value will become very large, i.e. the inverse of condition value will become near 0;

2) $h_1(\boldsymbol{J}_i, \boldsymbol{J}_j) = \frac{1}{2}\|\boldsymbol{J}_i - \boldsymbol{J}_j\|^2$

This index is used to judge whether two hidden units $\boldsymbol{J}_i$ and $\boldsymbol{J}_j$ overlap. If MLPs has been affected by overlap singularity, $\boldsymbol{J}_i = \boldsymbol{J}_j$, then $h_1(\boldsymbol{J}_i, \boldsymbol{J}_j) = 0$;

3) $h_2(\boldsymbol{J}_i, \boldsymbol{J}_j) = \frac{1}{2}\|\boldsymbol{J}_i + \boldsymbol{J}_j\|^2$

This index is used to judge whether MLPs have been affected by opposite singularity. If MLPs has been affected by opposite singularity, $\boldsymbol{J}_i = -\boldsymbol{J}_j$, then $h_2(\boldsymbol{J}_i, \boldsymbol{J}_j) = 0$.

Then we will take two experiments to visually represent the learning dynamics of MLPs, which will verify the correctness of Theorem 1 and illustrate the potential to design better algorithms based on the obtained analytical form of FIM.

For given teacher parameters, by choosing the initial student parameters, we use gradient descent method to accomplish the training processes. In the following figures of experiment results, '$\circ$' and '$\times$' represent the initial state and final state, respectively.

A. LEARNING TRAJECTORIES IN ERROR FUNCTION BASED MLPs

This experiment is taken to verify the correctness of the obtained analytical form of FIM, i.e. on the singularity, the FIM is singular and otherwise the FIM is regular. We choose the teacher and student model to both have 6 hidden units, i.e. $M = 6$ and $k = 6$. The additional noise is $\varepsilon \sim N(0, 0.05)$ and the covariance matrix of Gaussian input is $\boldsymbol{\Sigma} = \begin{bmatrix} 0.8 & 0.3 \\ 0.3 & 0.6 \end{bmatrix}$. The learning rate is chosen as $\eta = 0.002$. Then we give the singular cases of learning dynamics which are affected by singularities and regular case, respectively.

Case 1 (Opposite Singularity): the learning process is influenced by opposite singularity.

In this case, the learning process is affected by opposite singularity. We choose the teacher parameters are:

$$\boldsymbol{t} = [\boldsymbol{t}_1, \boldsymbol{t}_2, \boldsymbol{t}_3, \cdots, \boldsymbol{t}_6] = \begin{bmatrix} -1.5755 & -1.5637 & -0.5704 & -0.1654 & 0.6669 & 1.8897 \\ 0.6475 & -1.8524 & 0.1433 & -0.6730 & 1.9557 & 1.0012 \end{bmatrix}, \quad (35)$$

$$\boldsymbol{v} = [v_1, v_2, v_3, \cdots, v_6] = [1.3678,\ 1.3952,\ 0.3849,\ -0.8077,\ 1.3364,\ -1.0324]. \quad (36)$$

The initial student parameters are:

$$\boldsymbol{J}^{(0)} = \left[\boldsymbol{J}_1^{(0)}, \boldsymbol{J}_2^{(0)}, \boldsymbol{J}_3^{(0)}, \cdots, \boldsymbol{J}_6^{(0)}\right] = \begin{bmatrix} -1.6520 & -1.1852 & -0.9653 & 1.9594 & -0.4168 & -0.0718 \\ -1.6410 & -0.2991 & 1.9378 & -1.4052 & 0.4970 & -1.4433 \end{bmatrix}, \quad (37)$$

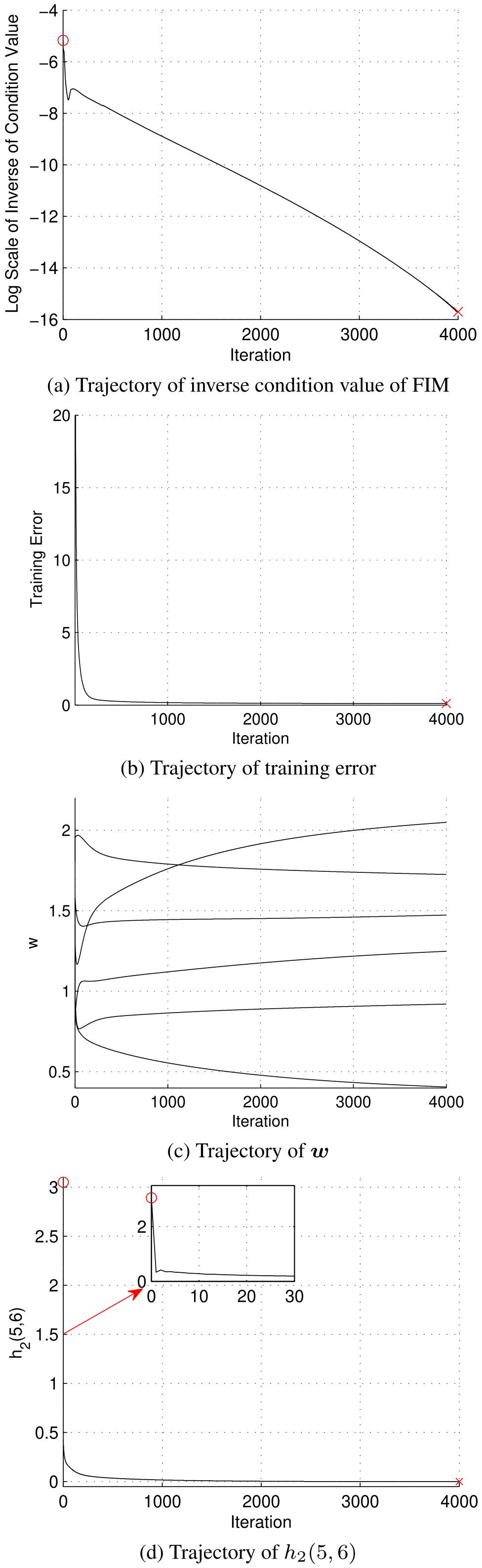
$$\boldsymbol{w}^{(0)} = [w_1^{(0)}, w_2^{(0)}, w_3^{(0)}, w_4^{(0)}, w_5^{(0)}, w_6^{(0)}] = [1.4240,\ 0.7876,\ 1.2879,\ 1.2908,\ 1.8043,\ 1.2281]. \quad (38)$$

The final student parameters are:

$$\boldsymbol{J} = [\boldsymbol{J}_1, \boldsymbol{J}_2, \boldsymbol{J}_3, \cdots, \boldsymbol{J}_6] = \begin{bmatrix} -1.7850 & -1.6606 & 0.3827 & 0.8872 & -0.7958 & 0.7886 \\ -1.3509 & 0.6381 & 1.5688 & -1.4834 & 0.9384 & -0.9560 \end{bmatrix}, \quad (39)$$

$$\boldsymbol{w} = [w_1, w_2, w_3, w_4, w_5, w_6] = [2.0482,\ 1.2480,\ 1.4727,\ 0.4057,\ 1.7249,\ 0.9203]. \quad (40)$$

The experiment results are shown in Fig. 2, which represent the trajectories of log scale of inverse condition number of FIM, training error, output weights $\boldsymbol{w}$ and $h_2(5, 6)$, respectively.



(a) Trajectory of inverse condition value of FIM

(b) Trajectory of training error

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

(d) Trajectory of $h_2(5, 6)$

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

From Fig. 2(d), it can be seen that $h_2(5, 6)$ fast becomes nearly 0 when the training process has started. When the training finishes, as shown in Eq. (39) which is the final state of student parameters, hidden units $\boldsymbol{J}_5$ and $\boldsymbol{J}_6$ are nearly opposite. The learning process is affected by opposite singularity. Meanwhile, as can be seen in Fig. 2(a), the inverse condition value of FIM is smaller than $10E-15$ till the end of the training process, which implies FIM becomes nearly singular. This is in accordance with theoretical analysis.

Case 2 (Overlap Singularity): the learning process is affected by the overlap singularity.

For this case, two hidden units overlap during the learning process and the learning dynamics are trapped in the overlap singularity. We choose the teacher parameters are:

$$\boldsymbol{t} = \begin{bmatrix} 1.5735 & 1.4947 & -0.6714 & -1.4781 & -0.8529 & 1.8804 \\ 0.4842 & -0.3383 & 0.4107 & -1.5900 & 1.2991 & -1.8494 \end{bmatrix}, \quad (41)$$

$$\boldsymbol{v} = [1.1691,\ 0.2711,\ -0.9127,\ -0.0828,\ -1.0184,\ -0.8792]. \quad (42)$$

The initial student parameters are:

$$\boldsymbol{J}^{(0)} = \begin{bmatrix} -0.4005 & -0.7154 & 0.5955 & 1.8786 & -1.7056 & 1.7659 \\ -1.9145 & -0.1857 & -0.8117 & 0.0666 & -1.4145 & -1.6531 \end{bmatrix}, \quad (43)$$

$$\boldsymbol{w}^{(0)} = [-1.9906,\ 1.8752,\ 0.0548,\ 1.4312,\ 1.3508,\ -0.8365]. \quad (44)$$

The final student parameters are:

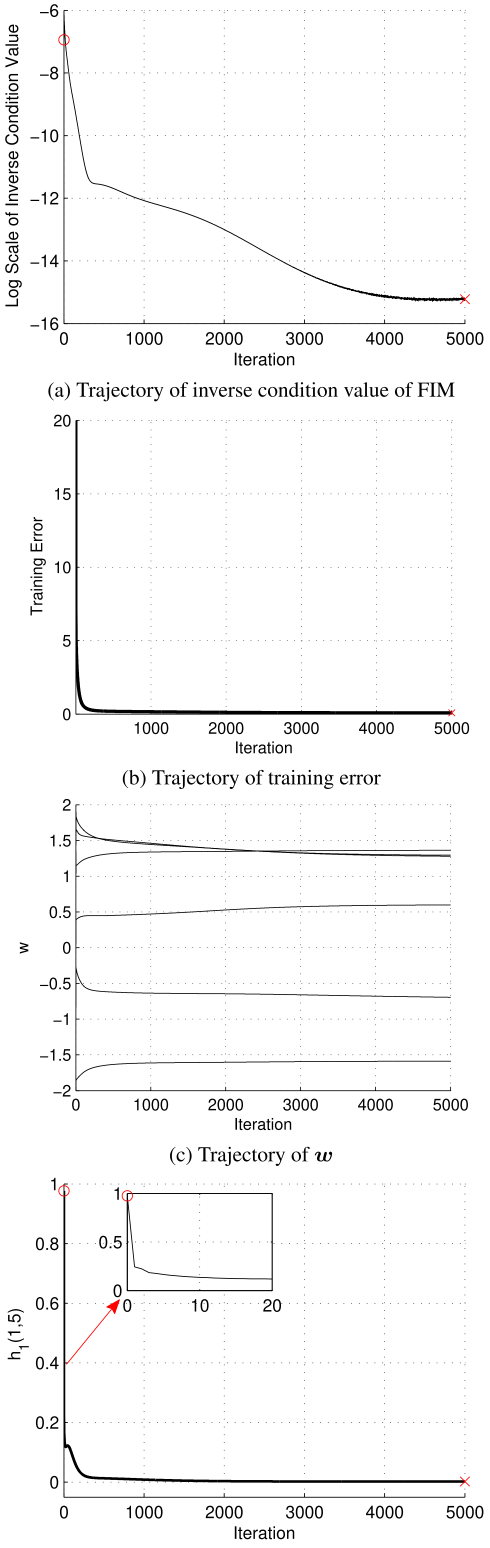
$$\boldsymbol{J} = \begin{bmatrix} -1.2387 & 0.6071 & 0.7770 & 1.6056 & -1.2290 & 1.9885 \\ -1.5856 & -0.4755 & -1.4142 & 0.2559 & -1.6032 & -1.9087 \end{bmatrix}, \quad (45)$$

$$\boldsymbol{w} = [-1.5883,\ 1.2941,\ 0.5979,\ 1.2793,\ 1.3623,\ -0.6943]. \quad (46)$$

The experiment results are shown in Fig. 3, which represent the trajectories of log scale of inverse condition number of FIM, training error, output weights $\boldsymbol{w}$ and $h_1(1,5)$, respectively.

From Fig. 3(d) and the final states of student parameters, we can see that $\boldsymbol{J}_1$ and $\boldsymbol{J}_5$ nearly overlap, which implies that the learning process is affected by overlap singularity. As also can be seen in Fig. 3(a), the inverse condition value of FIM decrease fast to nearly 0 and is finally smaller than $10E-15$, thus the FIM becomes nearly singular till the end when the learning process has been affected by overlap singularity.

Remark 1: It can be seen that the log scale of the inverse of condition value obviously fluctuates at the end of the learning process (Figure 3(a)). We think this is mainly because the value is too small (smaller than $10E-15$), and even a slight change of the parameters would cause the obvious fluctuation of the condition number of the Fisher information matrix due to the limit to the degree of accuracy of computer.



(a) Trajectory of inverse condition value of FIM

(b) Trajectory of training error

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

(d) Trajectory of $h_1(1,5)$

FIGURE 3. **Case 2 (Overlap singularity) in error function based MLPs.**

Case 3 (Elimination Singularity): the learning process is affected by the elimination singularity.

For this case, one output weight crosses 0 during the learning process and a plateau phenomenon can be obviously observed. We choose the teacher parameters are:

$$\boldsymbol{t} = \begin{bmatrix} -1.1997 & -1.0310 & -0.1054 & 1.2778 & 1.8295 & 1.9685 \\ -0.3513 & 0.5353 & 1.5588 & 1.9941 & 0.2635 & 0.9331 \end{bmatrix}, \tag{47}$$

$$\boldsymbol{v} = [-0.2133,\ 0.3684,\ -1.1383,\ -0.6795,\ 1.8381,\ 1.3734]. \tag{48}$$

The initial student parameters are:

$$\boldsymbol{J}^{(0)} = \begin{bmatrix} -0.9311 & -1.4760 & 1.5680 & -1.9608 & -0.5129 & 1.2863 \\ 1.9021 & -0.6705 & -0.4289 & 0.0666 & -1.4145 & -1.6531 \end{bmatrix}, \tag{49}$$

$$\boldsymbol{w}^{(0)} = [-1.2450,\ 0.4280,\ -1.0404,\ 0.2285,\ -0.3309,\ -0.9585]. \tag{50}$$

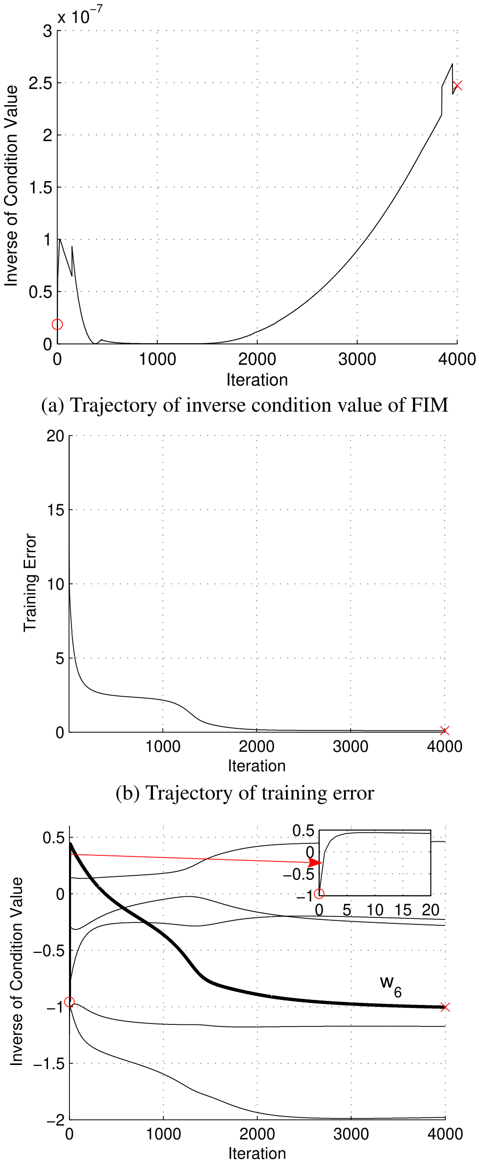
The final student parameters are:

$$\boldsymbol{J} = \begin{bmatrix} -2.1202 & -1.7697 & 1.2750 & -1.5695 & -1.4591 & 0.0598 \\ 0.9605 & -1.5986 & -0.6385 & 2.2159 & 0.2291 & 1.6207 \end{bmatrix}, \tag{51}$$

$$\boldsymbol{w} = [-1.1747,\ -1.9794,\ -0.2814,\ 0.4621,\ -0.2280,\ -1.0043]. \tag{52}$$

The experiment results are shown in Fig. 4, which represent the trajectories of inverse condition number of FIM, training error, and output weights $\boldsymbol{w}$, respectively.

From Fig. 4(c), we can see that w_6 crosses 0 in the learning process and the learning process is affected by elimination singularity. During the stage w_6 crosses 0, the plateau phenomenon can be obviously observed in Fig. 4(b), and FIM also degenerates at this stage (Fig. 4(a)). Then the student parameters escape the influence of elimination singularity and finally converge to the global minimum which can be seen from the final state of student parameters (51)-(52), meanwhile, the FIM also becomes regular in the late stage in Fig. 4(a) as the learning dynamics are not influenced by elimination singularity.



(a) Trajectory of inverse condition value of FIM

(b) Trajectory of training error

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

FIGURE 4. Case 3 (Elimination singularity) in error function based MLPs.

Case 4 (Fast Convergence): the learning process does not suffer from the influence of singularities

For this case, the learning dynamics are not influenced by any singularity and fast converge to the optimal value. we choose the teacher parameters are:

$$\boldsymbol{t} = \begin{bmatrix} -1.7244 & -1.9571 & -0.2937 & -0.6621 & 0.5371 & 1.2059 \\ 0.6628 & -1.9488 & -0.6609 & 1.9513 & 1.6839 & 0.2742 \end{bmatrix}, \tag{53}$$

$$\boldsymbol{v} = [1.6296,\ -1.2374,\ 1.7883,\ 1.7996,\ -0.5046,\ -1.0187]. \tag{54}$$

The initial student parameters are:

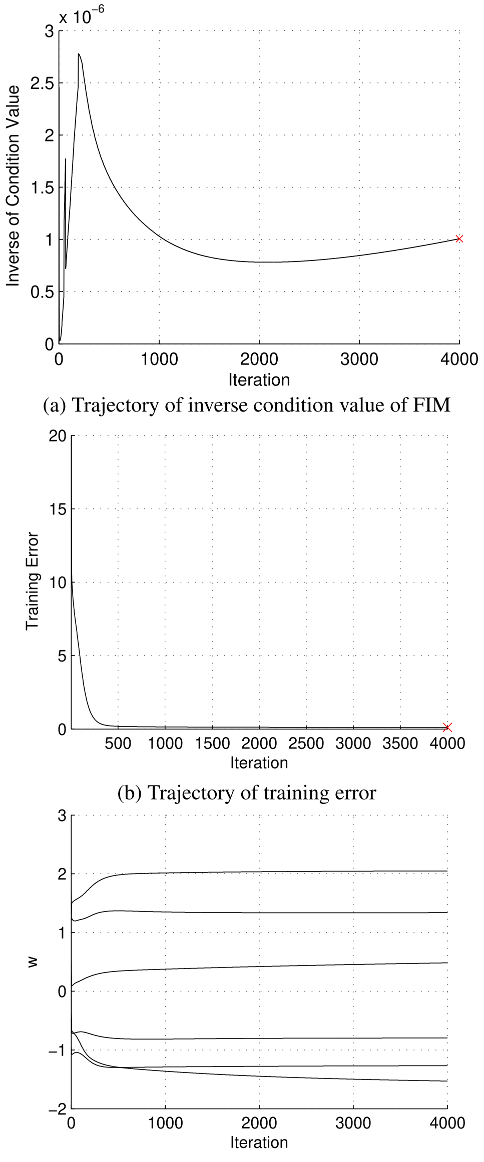
$$\boldsymbol{J}^{(0)} = \begin{bmatrix} -1.8579 & -0.5237 & -0.5872 & -1.3052 & 1.8091 & 0.7872 \\ -1.2815 & -0.8620 & -1.8846 & -0.0195 & -1.0615 & -1.0158 \end{bmatrix}, \tag{55}$$

$$\boldsymbol{w}^{(0)} = [1.6435,\ 1.7157,\ 1.3399,\ -1.0196,\ 0.4613,\ 0.5325]. \tag{56}$$

The final student parameters are:

$$\boldsymbol{J} = \begin{bmatrix} -1.6574 & -0.4919 & -0.7439 & -1.9050 & 1.6764 & 0.7410 \\ 0.2770 & -0.5973 & -1.3315 & -1.8704 & -0.8679 & -2.0478 \end{bmatrix}, \tag{57}$$

$$\boldsymbol{w} = [1.3381,\ 2.0474,\ 0.4835,\ -1.2688,\ -0.7970,\ -1.5295]. \tag{58}$$



(a) Trajectory of inverse condition value of FIM

(b) Trajectory of training error

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

FIGURE 5. **Case 4 (Fast convergence) in error function based MLPs.**

The experiment results are shown in Fig. 5, which represent the trajectories of the inverse condition number of FIM, training error and output weights $\boldsymbol{w}$, respectively.

As can be seen from Fig. 5(b) and the final student parameters, the learning dynamics quickly converge to the global minimum and have not been affected by any singularity. The FIM also remains regular during the entire training process.

In the above 4 cases, we have shown the learning dynamics belong to singular cases and regular case, respectively. We can see that the FIM degenerates when the learning dynamics are affected by singularities and remains regular in other cases, which verifies the correctness of the obtained results in Theorem 1.

Remark 2: Compared with bipolar error function, hyperbolic tangent function $tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ is the most widely used bipolar activation function in MLPs. Although the theoretical results in Theorem 1 are obtained based on bipolar error function, we take another experiment to illustrate that the results are also valid for hyperbolic tangent function based MLPs. The experiment set up is the same as in section 4.1. We choose the teacher parameters and initial student parameters just the same in section 4.1. The only difference is that hyperbolic tangent function is used to replace the bipolar error function as the activation function in the teacher and student models. The experiment results are basically the same with the results shown in section 4.1. Thus the analytical form of the FIM based on bipolar error function can also be applied to the hyperbolic tangent function based MLPs.

B. FIM BASED NATURAL GRADIENT DESCENT ALGORITHM

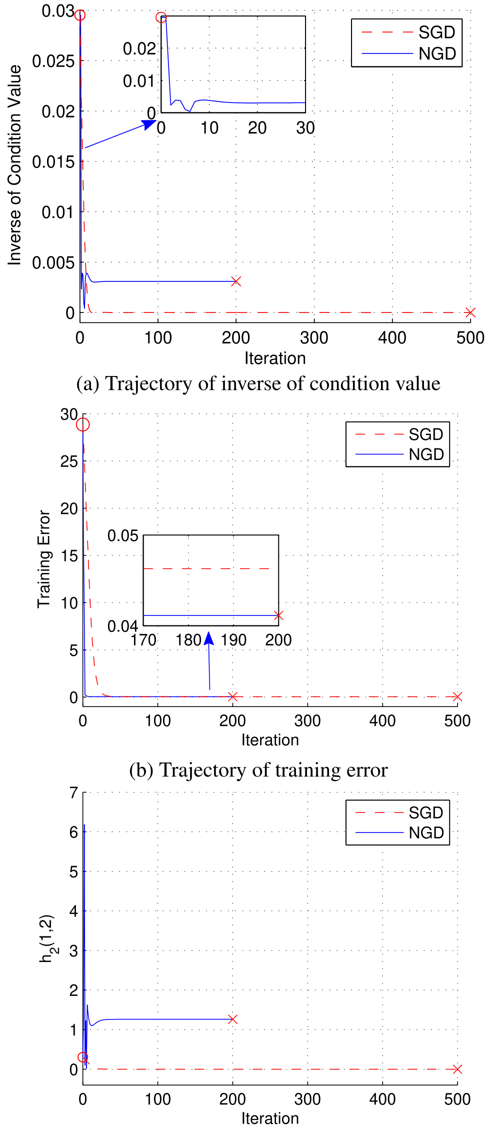
As the natural gradient descent direction becomes the steepest descent direction, researchers proposed natural gradient method to overcome the influence of singularities, the parameter modification formula is shown as follow:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \boldsymbol{F}(\boldsymbol{\theta}_t)^{-1} \frac{\partial l(y_t, \boldsymbol{x}_t, \boldsymbol{\theta}_t)}{\partial \boldsymbol{\theta}_t}, \tag{59}$$

where η is the learning rate and $\boldsymbol{F}(\boldsymbol{\theta}_t)$ is the FIM at iteration t. Compared to standard gradient descent method, the natural gradient method adds the inverse FIM item to the modification of parameters.

From (59), we can see that computing the inverse FIM plays a key role in natural gradient descent method. Unfortunately, it is very hard to obtain the analytical form of inverse FIM and directly computing the inverse FIM also requires enormous computation cost. This limits the application of natural gradient descent method. Then researchers proposed adaptive natural gradient descent method, which used an iteration formula to approximate the inverse FIM instead of directly computing it. Although computing the inverse of large dimension matrix still faces many difficulties, the analytical form of FIM can help us to investigate better approximation formula of inverse FIM, which will lead to a significant improvement of adaptive natural gradient descent algorithms.

In this experiment, we aim to present the performance of natural gradient method by directly computing the inverse FIM based on the obtained analytical form. We choose the teacher model and student model both have 2 hidden nodes and the input dimension is 1, i.e. $M = 2$, $k = 2$ and $n = 1$. Then the experiment results will be shown by comparing natural gradient descent (NGD) algorithm with standard gradient descent (SGD) algorithm, where three singular cases, including opposite singularity case, overlap singularity case and elimination singularity case, are investigated. Due to the difficulty in calculating the inverse FIM and the precision limitation of the computer, we set the initial state of part of student parameters to be optimal value, and only the rest part of student parameters need to be modified.

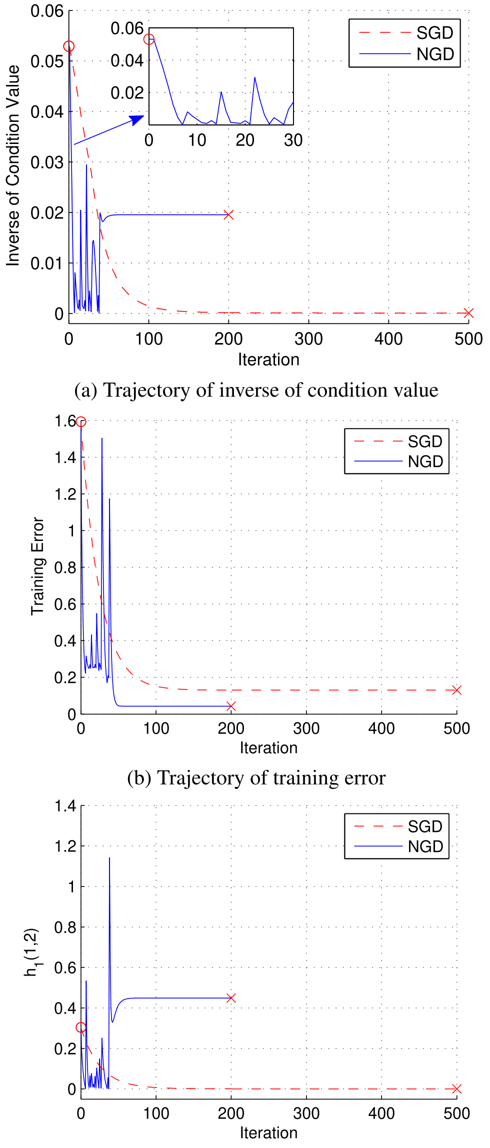


(a) Trajectory of inverse of condition value

(b) Trajectory of training error

(c) Trajectory of J

FIGURE 6. **Case 1 (Opposite singularity) with NGD algorithm and SGD algorithm (The final state of student parameters using SGD algorithm are $J_1 = -0.2022$ and $J_2 = 0.1917$. The final state of student parameters of NGD algorithm are $J_1 = -0.4945$ and $J_2 = -1.0379$.)**



(a) Trajectory of inverse of condition value

(b) Trajectory of training error

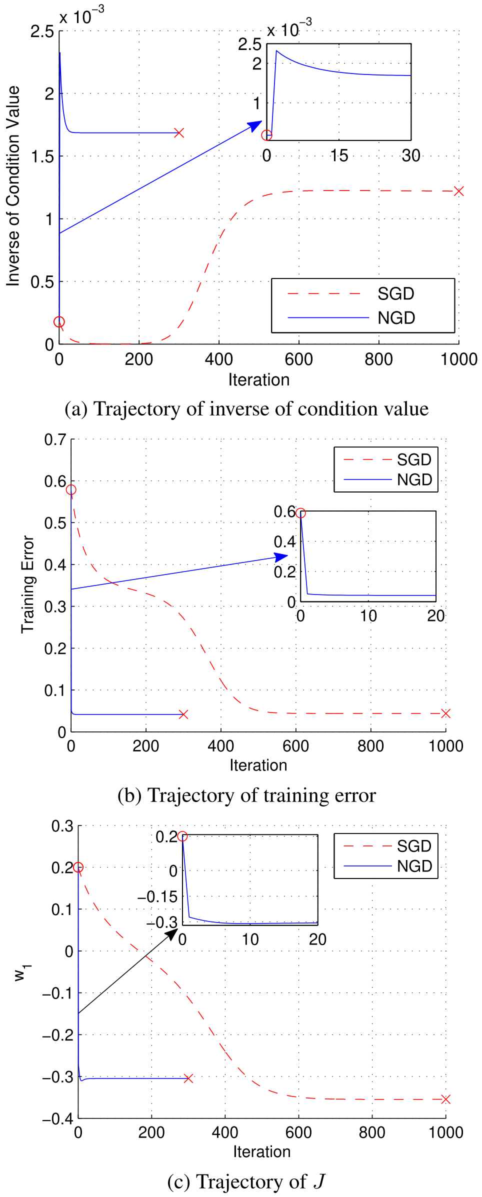
(c) Trajectory of J

FIGURE 7. **Case 2 (Overlap singularity) with NGD algorithm and SGD algorithm (The final state of student parameters using SGD algorithm are $J_1 = -0.6318$ and $J_2 = -0.6024$. The final state of student parameters using NGD algorithm are $J_1 = 0.4436$ and $J_2 = 1.3913$.)**

Case 1 (Opposite Singularity): For this case, we only modify the student parameter J_1 and J_2, w_1 and w_2 are fixed to be the optimal value and remains unchanged, i.e. $w_1 = v_1$ and $w_2 = v_2$. We choose the teacher parameters are: $t_1 = -0.49$, $t_2 = -1.00$, $v_1 = 0.95$, and $v_2 = -0.25$. The initial state of student parameters are: $J_1^{(0)} = 0.98$, $J_2^{(0)} = -0.20$. The MLP is trained 200 epochs using NGD algorithm and 500 epochs using SGD algorithm, respectively. The experiment results are shown in Fig. 6, which represent the trajectories of inverse condition number of FIM, training error, and $h_2(1, 2)$, respectively.

Case 2 (Overlap Singularity): For this case, w_1 and w_2 are set to be the optimal value and we only modify the student parameter J_1 and J_2. We choose the teacher parameters are: $t_1 = 0.45$, $t_2 = 1.38$, $v_1 = -0.56$, and $v_2 = 0.37$. The initial state of student parameters are: $J_1^{(0)} = -0.99$, $J_2^{(0)} = -0.21$. The MLP is trained 200 epochs using NGD algorithm and 500 epochs using SGD algorithm, respectively. The experiment results are shown in Fig. 11, which represent the trajectories of inverse condition number of FIM, training error, and $h_1(1, 2)$, respectively.

Case 3 (Elimination Singularity): For this case, J_2 and w_2 remain invariable in the training process, i.e. $J_2 = t_2$ and $w_2 = v_2$. Only the student parameters J_1 and w_1 are needed to be modified. We choose the teacher parameters are: $t_1 = 0.40$, $t_2 = 0.89$, $v_1 = -0.32$, and $v_2 = -0.90$. The initial state of student parameters are: $J_1^{(0)} = 0.21$, $w_1^{(0)} = 0.20$. The MLP is trained 300 epochs using NGD algorithm and 1000 epochs using SGD algorithm, respectively. The



(a) Trajectory of inverse of condition value

(b) Trajectory of training error

(c) Trajectory of J

FIGURE 8. Case 3 (Elimination singularity) with NGD algorithm and SGD algorithm (The final state of student parameters using SGD algorithm are $J_1 = 0.3606$ and $w_1 = -0.3547$. The final state of student parameters using NGD algorithm are $J_1 = 0.4255$ and $w_1 = -0.3049$.)

experiment results are shown in Fig. 11, which represent the trajectories of inverse condition number of FIM, training error, and w_1, respectively.

For case 1 and case 2, as can be seen from Fig. 6 - Fig. 7, when using SGD algorithm to train the MLPs, the student parameters are trapped in opposite singularity or overlap singularity till the end. In sharp contrast, when using NGD algorithm to train the MLPs, the learning dynamics can easily escape the influence of opposite singularity and overlap singularity and converge to the global minimum. For case 3, from Fig. 8, we can see that the learning process is affected by elimination singularity. A plateau phenomenon can be observed in Fig. 8(b) for SGD algorithm case and the natural gradient algorithm can significantly reduce the influence of elimination singularity.

All the experiment results of above three cases have illustrated the efficiency of FIM based natural gradient method to overcome the influence of singularities. Since the analytical form of FIM have obtained in Theorem 1, it is important to derive better approximation formula of inverse FIM based on Theorem 1 in the future, which will facilitate the application of natural gradient method to high-dimensional systems.

V. CONCLUSION AND DISCUSSIONS

Multilayer perceptrons have been widely used in many field, however the singularities existed in the parameter space often seriously influence the learning dynamics. As Fisher information matrix degenerates on the singularities, the FIM plays a significant role in investigating the singular learning dynamics. In this paper, for MLPs with general Gaussian input, by choosing the bipolar error function as the activation function, we obtain the analytical form of FIM. In the experiment part, we have verified the correctness of the analytical form, and finally showed the efficiency of FIM-based NGD algorithm in comparison with SGD algorithm. In the future, based on the obtained analytical form of FIM, we aim to derive better approximation formulas of inverse FIM that can be applied to high-dimensional systems.

APPENDIX

THE ANALYTICAL FORM OF $Q_1(J_i, J_j)$, $Q_2(J_i, J_j)$ AND $Q_3(J_i, J_j)$

From Eq. (2), we have

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

then

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

$Q_1(\boldsymbol{J}_i, \boldsymbol{J}_j)$, $Q_2(\boldsymbol{J}_i, \boldsymbol{J}_j)$ and $Q_3(\boldsymbol{J}_i, \boldsymbol{J}_j)$ can be rewritten as:

$$\begin{aligned}
&Q_1(\boldsymbol{J}_i, \boldsymbol{J}_j)\\
&= \left(\sqrt{2\pi}\right)^{-n}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\frac{\partial\phi(\boldsymbol{x},\boldsymbol{J}_i)}{\partial\boldsymbol{J}_i}\frac{\partial\phi(\boldsymbol{x},\boldsymbol{J}_j)}{\partial\boldsymbol{J}_j^T}\\
&\quad\times\exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right)\\
&\quad\times\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(y-f_0(\boldsymbol{x}))^2\right)\mathrm{d}y\mathrm{d}\boldsymbol{x}\\
&= \left(\sqrt{2\pi}\right)^{-n}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\int_{-\infty}^{+\infty}\frac{\partial\phi(\boldsymbol{x},\boldsymbol{J}_i)}{\partial\boldsymbol{J}_i}\frac{\partial\phi(\boldsymbol{x},\boldsymbol{J}_j)}{\partial\boldsymbol{J}_j^T}\\
&\quad\times\exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x}.
\end{aligned}\tag{A-3}$$

$$\begin{aligned}
Q_2(\boldsymbol{J}_i, \boldsymbol{J}_j) &= \left(\sqrt{2\pi}\right)^{-n}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\phi(\boldsymbol{x},\boldsymbol{J}_j)\\
&\quad\times\frac{\partial\phi(\boldsymbol{x},\boldsymbol{J}_i)}{\partial\boldsymbol{J}_i}\exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right)
\end{aligned}$$

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

$$
\begin{aligned}
Q_3(\boldsymbol{J}_i,\boldsymbol{J}_j) &= \left(\sqrt{2\pi}\right)^{-n} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \phi(\boldsymbol{x},\boldsymbol{J}_i)\\
&\quad \times \phi(\boldsymbol{x},\boldsymbol{J}_j) \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right)\\
&\quad \times \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y-f_0(\boldsymbol{x}))^2}{2}\right)\mathrm{d}y\mathrm{d}\boldsymbol{x}\\
&= \left(\sqrt{2\pi}\right)^{-n} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \int_{-\infty}^{+\infty} \phi(\boldsymbol{x},\boldsymbol{J}_i)\\
&\quad \times \phi(\boldsymbol{x},\boldsymbol{J}_j) \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x}.
\end{aligned}
\tag{A-5}
$$

We denote $\boldsymbol{A}(\boldsymbol{J}_i,\boldsymbol{J}_j)^{-1} = \boldsymbol{\Sigma}^{-1} + \boldsymbol{J}_i\boldsymbol{J}_i^T + \boldsymbol{J}_j\boldsymbol{J}_j^T$ and $\boldsymbol{B}(\boldsymbol{J}_i) = \boldsymbol{\Sigma}^{-1} + \boldsymbol{J}_i\boldsymbol{J}_i^T$, then we can have:

$$
\begin{aligned}
\boldsymbol{Q}_1(\boldsymbol{J}_i,\boldsymbol{J}_j) &= \left(\sqrt{2\pi}\right)^{-n} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \int_{-\infty}^{+\infty} \frac{\partial \phi(\boldsymbol{x},\boldsymbol{J}_i)}{\partial \boldsymbol{J}_i}\\
&\quad \times \frac{\partial \phi(\boldsymbol{x},\boldsymbol{J}_j)}{\partial \boldsymbol{J}_j^T} \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x}\\
&= \frac{2}{\pi}\left(\sqrt{2\pi}\right)^{-n} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \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)\\
&\quad \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{\Sigma}^{-1}\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x}\\
&= \frac{2}{\pi}\left(\sqrt{2\pi}\right)^{-n} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \int_{-\infty}^{+\infty} \boldsymbol{x}\boldsymbol{x}^T\\
&\quad \times \exp\left(-\frac{1}{2}\boldsymbol{x}^T(\boldsymbol{\Sigma}^{-1} + \boldsymbol{J}_i\boldsymbol{J}_i^T + \boldsymbol{J}_j\boldsymbol{J}_j^T)\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x}\\
&= \frac{2}{\pi}\left(\sqrt{2\pi}\right)^{-n} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \int_{-\infty}^{+\infty} \boldsymbol{x}\boldsymbol{x}^T\\
&\quad \times \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}(\boldsymbol{J}_i,\boldsymbol{J}_j)^{-1}\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x}\\
&= \frac{2}{\pi}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}|\boldsymbol{A}(\boldsymbol{J}_i,\boldsymbol{J}_j)|^{\frac{1}{2}}\left(\sqrt{2\pi}\right)^{-n}|\boldsymbol{A}(\boldsymbol{J}_i,\boldsymbol{J}_j)|^{-\frac{1}{2}}\\
&\quad \times \int_{-\infty}^{+\infty} \boldsymbol{x}\boldsymbol{x}^T \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}(\boldsymbol{J}_i,\boldsymbol{J}_j)^{-1}\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x}\\
&= \frac{2}{\pi}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}|\boldsymbol{A}(\boldsymbol{J}_i,\boldsymbol{J}_j)|^{\frac{1}{2}}\boldsymbol{A}(\boldsymbol{J}_i,\boldsymbol{J}_j),
\end{aligned}
\tag{A-6}
$$

$$
\begin{aligned}
Q_2(\boldsymbol{J}_i,\boldsymbol{J}_j) &= \left(\sqrt{2\pi}\right)^{-n} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \int_{-\infty}^{+\infty} \frac{\partial \phi(\boldsymbol{x},\boldsymbol{J}_i)}{\partial \boldsymbol{J}_i}\\
&\quad \times \phi(\boldsymbol{x},\boldsymbol{J}_j) \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x}\\
&= \sqrt{\frac{2}{\pi}}\left(\sqrt{2\pi}\right)^{-n} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \int_{-\infty}^{+\infty} \boldsymbol{x}\exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{J}_i\boldsymbol{J}_i^T\boldsymbol{x}\right)\\
&\quad \times \phi(\boldsymbol{x},\boldsymbol{J}_j) \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x}\\
&= \sqrt{\frac{2}{\pi}}\left(\sqrt{2\pi}\right)^{-n} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \int_{-\infty}^{+\infty} \boldsymbol{x}\phi(\boldsymbol{x},\boldsymbol{J}_j)\\
&\quad \times \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{B}(\boldsymbol{J}_i)\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x}\\
&= \sqrt{\frac{2}{\pi}}\left(\sqrt{2\pi}\right)^{-n} |\boldsymbol{\Sigma}|^{-\frac{1}{2}}\boldsymbol{B}(\boldsymbol{J}_i)^{-1}\\
&\quad \times \int_{-\infty}^{+\infty} \phi(\boldsymbol{x},\boldsymbol{J}_j)\mathrm{d}\exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{B}(\boldsymbol{J}_i)\boldsymbol{x}\right)\\
&= \sqrt{\frac{2}{\pi}}\left(\sqrt{2\pi}\right)^{-n} |\boldsymbol{\Sigma}|^{-\frac{1}{2}}\boldsymbol{B}(\boldsymbol{J}_i)^{-1}\\
&\quad \times\phi(\boldsymbol{x},\boldsymbol{J}_j)\exp\left(\frac{1}{2}\boldsymbol{x}^T\boldsymbol{B}(\boldsymbol{J}_i)\boldsymbol{x}\right)\Big|_{-\infty}^{+\infty}\\
&\quad + \frac{2}{\pi}\left(\sqrt{2\pi}\right)^{-n} |\boldsymbol{\Sigma}|^{-\frac{1}{2}}\boldsymbol{B}(\boldsymbol{J}_i)^{-1}\boldsymbol{J}_j\\
&\quad \times\int_{-\infty}^{+\infty}\exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{J}_j\boldsymbol{J}_j^T\boldsymbol{x}\right)\\
&\quad \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{B}(\boldsymbol{J}_i)\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x}\\
&= \frac{2}{\pi}\left(\sqrt{2\pi}\right)^{-n} |\boldsymbol{\Sigma}|^{-\frac{1}{2}}\boldsymbol{B}(\boldsymbol{J}_i)^{-1}\boldsymbol{J}_j\\
&\quad \times\int_{-\infty}^{+\infty}\exp\left(-\frac{1}{2}\boldsymbol{x}^T(\boldsymbol{B}(\boldsymbol{J}_i)+\boldsymbol{J}_j\boldsymbol{J}_j^T)\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x}\\
&= \frac{2}{\pi}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\boldsymbol{B}(\boldsymbol{J}_i)^{-1}\boldsymbol{J}_j\left(\sqrt{2\pi}\right)^{-n}\\
&\quad \times\int_{-\infty}^{+\infty}\exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}(\boldsymbol{J}_i,\boldsymbol{J}_j)^{-1}\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x}\\
&= \frac{2}{\pi}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}|\boldsymbol{A}(\boldsymbol{J}_i,\boldsymbol{J}_j)|^{\frac{1}{2}}\boldsymbol{B}(\boldsymbol{J}_i)^{-1}\boldsymbol{J}_j,
\end{aligned}
\tag{A-7}
$$

where:

$$
\begin{aligned}
\boldsymbol{B}(\boldsymbol{J}_i)^{-1} &= \left(\boldsymbol{\Sigma}^{-1}+\boldsymbol{J}_i\boldsymbol{J}_i^T\right)^{-1}\\
&= \boldsymbol{\Sigma} - \frac{\boldsymbol{\Sigma}\boldsymbol{J}_i\boldsymbol{J}_i^T\boldsymbol{\Sigma}}{1+\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_i}.
\end{aligned}
\tag{A-8}
$$

$$
\begin{aligned}
\boldsymbol{A}(\boldsymbol{J}_i,\boldsymbol{J}_j) &= \left(\boldsymbol{B}(\boldsymbol{J}_i)+\boldsymbol{J}_j\boldsymbol{J}_j^T\right)^{-1}\\
&= \boldsymbol{B}(\boldsymbol{J}_i)^{-1} - \frac{\boldsymbol{B}(\boldsymbol{J}_i)^{-1}\boldsymbol{J}_j\boldsymbol{J}_j^T\boldsymbol{B}(\boldsymbol{J}_i)^{-1}}{1+\boldsymbol{J}_j^T\boldsymbol{B}(\boldsymbol{J}_i)^{-1}\boldsymbol{J}_j}\\
&= \boldsymbol{\Sigma} - \frac{\boldsymbol{\Sigma}\boldsymbol{J}_i\boldsymbol{J}_i^T\boldsymbol{\Sigma}}{1+\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_i}\\
&\quad - \frac{\left(\boldsymbol{\Sigma}-\frac{\boldsymbol{\Sigma}\boldsymbol{J}_i\boldsymbol{J}_i^T\boldsymbol{\Sigma}}{1+\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_i}\right)\boldsymbol{J}_j\boldsymbol{J}_j^T\left(\boldsymbol{\Sigma}-\frac{\boldsymbol{\Sigma}\boldsymbol{J}_i\boldsymbol{J}_i^T\boldsymbol{\Sigma}}{1+\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_i}\right)}{1+\boldsymbol{J}_j^T\left(\boldsymbol{\Sigma}-\frac{\boldsymbol{\Sigma}\boldsymbol{J}_i\boldsymbol{J}_i^T\boldsymbol{\Sigma}}{1+\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_i}\right)\boldsymbol{J}_j}\\
&= (\boldsymbol{\Sigma}^{-1}+\boldsymbol{J}_i\boldsymbol{J}_i^T+\boldsymbol{J}_j\boldsymbol{J}_j^T)^{-1} = \boldsymbol{\Sigma}-\boldsymbol{\Sigma}\\
&\quad \times\left(\frac{(1+\boldsymbol{J}_j^T\boldsymbol{\Sigma}\boldsymbol{J}_j)\boldsymbol{J}_i\boldsymbol{J}_i^T+(1+\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_i)\boldsymbol{J}_j\boldsymbol{J}_j^T}{(1+\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_i)(1+\boldsymbol{J}_j^T\boldsymbol{\Sigma}\boldsymbol{J}_j)-(\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_j)^2}\right.\\
&\quad \left.-\frac{\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_j(\boldsymbol{J}_i\boldsymbol{J}_j^T+\boldsymbol{J}_j\boldsymbol{J}_i^T)}{(1+\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_i)(1+\boldsymbol{J}_j^T\boldsymbol{\Sigma}\boldsymbol{J}_j)-(\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_j)^2}\right)\boldsymbol{\Sigma}.
\end{aligned}
\tag{A-9}
$$

According to the matrix determinant lemma:

$$|\boldsymbol{B}(\boldsymbol{J}_i)| = |\boldsymbol{\Sigma}^{-1} + \boldsymbol{J}_i\boldsymbol{J}_i^T| = (1+\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_i)|\boldsymbol{\Sigma}|^{-1}. \tag{A-10}$$

$$\begin{aligned}|\boldsymbol{A}(\boldsymbol{J}_i,\boldsymbol{J}_j)|^{-1} &= |\boldsymbol{B}(\boldsymbol{J}_i)+\boldsymbol{J}_j\boldsymbol{J}_j^T| = (1+\boldsymbol{J}_j^T\boldsymbol{B}(\boldsymbol{J}_i)^{-1}\boldsymbol{J}_j)|\boldsymbol{B}(\boldsymbol{J}_i)| \\ &= (1+\boldsymbol{J}_j^T\left(\boldsymbol{\Sigma}-\frac{\boldsymbol{\Sigma}\boldsymbol{J}_i\boldsymbol{J}_i^T\boldsymbol{\Sigma}}{1+\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_i}\right)(1+\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_i)|\boldsymbol{\Sigma}|^{-1} \\ &= ((1+\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_i)(1+\boldsymbol{J}_j^T\boldsymbol{\Sigma}\boldsymbol{J}_j) \\ &\quad -(\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_j)^2)|\boldsymbol{\Sigma}|^{-1}.\end{aligned} \tag{A-11}$$

$$|\boldsymbol{A}(\boldsymbol{J}_i,\boldsymbol{J}_j| = \frac{|\boldsymbol{\Sigma}|}{(1+\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_i)(1+\boldsymbol{J}_j^T\boldsymbol{\Sigma}\boldsymbol{J}_j)-(\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_j)^2}. \tag{A-12}$$

Then we calculate $Q_3(\boldsymbol{J}_i,\boldsymbol{J}_j)$.

$$\begin{aligned}Q_3(\boldsymbol{J}_i,\boldsymbol{J}_j) &= \int \boldsymbol{Q}_2(\boldsymbol{J}_i,\boldsymbol{J}_j)\,\mathrm{d}\boldsymbol{J}_i \\ &= \int \frac{2}{\pi}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}|\boldsymbol{A}(\boldsymbol{J}_i,\boldsymbol{J}_j)|^{\frac{1}{2}}\boldsymbol{B}(\boldsymbol{J}_i)^{-1}\boldsymbol{J}_j\,\mathrm{d}\boldsymbol{J}_i \\ &= \frac{2}{\pi}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\int \boldsymbol{B}(\boldsymbol{J}_i)^{-1}\boldsymbol{J}_j \\ &\quad\times\frac{|\boldsymbol{\Sigma}|^{\frac{1}{2}}}{\sqrt{(1+\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_i)(1+\boldsymbol{J}_j^T\boldsymbol{\Sigma}\boldsymbol{J}_j)-(\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_j)^2}}\,\mathrm{d}\boldsymbol{J}_i \\ &= \frac{2}{\pi}\int\frac{1}{\sqrt{1-\frac{(\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_j)^2}{(1+\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_i)(1+\boldsymbol{J}_j^T\boldsymbol{\Sigma}\boldsymbol{J}_j)}}} \\ &\quad\times\frac{\boldsymbol{B}(\boldsymbol{J}_i)^{-1}\boldsymbol{J}_j}{\sqrt{1+\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_i}\sqrt{1+\boldsymbol{J}_j^T\boldsymbol{\Sigma}\boldsymbol{J}_j}}\,\mathrm{d}\boldsymbol{J}_i \\ &= \frac{2}{\pi}\int\frac{1}{\sqrt{1-\frac{(\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_j)^2}{(1+\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_i)(1+\boldsymbol{J}_j^T\boldsymbol{\Sigma}\boldsymbol{J}_j)}}} \\ &\quad\times\mathrm{d}\frac{\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_j}{\sqrt{1+\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_i}\sqrt{1+\boldsymbol{J}_j^T\boldsymbol{\Sigma}\boldsymbol{J}_j}} \\ &= \frac{2}{\pi}\left(\arcsin\frac{\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_j}{\sqrt{1+\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_i}\sqrt{1+\boldsymbol{J}_j^T\boldsymbol{\Sigma}\boldsymbol{J}_j}}+C_0\right).\end{aligned} \tag{A-13}$$

As $Q_3(\boldsymbol{0},\boldsymbol{0}) = 0$, we can get $C_0 = 0$, then we have:

$$Q_3(\boldsymbol{J}_i,\boldsymbol{J}_j) = \frac{2}{\pi}\arcsin\frac{\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_j}{\sqrt{1+\boldsymbol{J}_i^T\boldsymbol{\Sigma}\boldsymbol{J}_i}\sqrt{1+\boldsymbol{J}_j^T\boldsymbol{\Sigma}\boldsymbol{J}_j}}. \tag{A-14}$$

□

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2014.

[3] L. Xie, W. Guo, H. Wei, Y. Tang, and D. Tao, "Efficient unsupervised dimension reduction for streaming multiview data," *IEEE Trans. Cybern.*, vol. 52, no. 3, pp. 1772–1784, Mar. 2022.

[4] S. Wen, X. Xie, Z. Yan, T. Huang, and Z. Zeng, "General memristor with applications in multilayer neural networks," *Neural Netw.*, vol. 103, pp. 142–149, Jul. 2018.

[5] G. Yang and H. Wang, "Multilayer neural network based asymptotic motion control of saturated uncertain robotic manipulators," *Appl. Intell.*, vol. 52, pp. 1–13, Jun. 2021.

[6] N. Dimitriou, L. Leontaris, T. Vafeiadis, D. Ioannidis, T. Wotherspoon, G. Tinker, and D. Tzovaras, "Fault diagnosis in microelectronics attachment via deep learning analysis of 3-D laser scans," *IEEE Trans. Ind. Electron.*, vol. 67, no. 7, pp. 5748–5757, Jul. 2020.

[7] A. Bibi, M. Alfadly, and B. Ghanem, "Analytic expressions for probabilistic moments of PL-DNN with Gaussian input," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 9099–9107.

[8] A. Ananthakrishnan and M. G. Allen, "All-passive hardware implementation of multilayer perceptron classifiers," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 9, pp. 4086–4095, Sep. 2021.

[9] G. Karimi and M. Heidarian, "Facial expression recognition with polynomial legendre and partial connection MLP," *Neurocomputing*, vol. 434, pp. 33–44, Apr. 2021.

[10] S. Watanabe, "Almost all learning machines are singular," in *Proc. IEEE Symp. Found. Comput. Intell.*, Honolulu, HI, USA, Apr. 2007, pp. 383–388.

[11] S.-I. Amari, H. Park, and T. Ozeki, "Singularities affect dynamics of learning in neuromanifolds," *Neural Comput.*, vol. 18, no. 5, pp. 1007–1065, 2006.

[12] H. Wei, J. Zhang, F. Cousseau, T. Ozeki, and S.-I. Amari, "Dynamics of learning near singularities in layered networks," *Neural Comput.*, vol. 20, no. 3, pp. 813–843, Mar. 2008.

[13] W. Guo, Y. Yang, Y. Zhou, Y. Tan, H. Wei, A. Song, and G. Pang, "Influence area of overlap singularity in multilayer perceptrons," *IEEE Access*, vol. 6, pp. 60214–60223, 2018.

[14] S.-I. Amari and T. Ozeki, "Differential and algebraic geometry of multilayer perceptrons," *IEICE Trans. Fundamentals Electron.*, vol. 84, no. 1, pp. 31–38, 2001.

[15] S.-I. Amari, "Neural learning in structured parameter spaces—Natural Riemannian gradient," in *Advances in Neural Information Processing Systems*, vol. 9, M. Mozer, M. I. Jordan, T. Petsche, Eds. Denver, CO, USA: MIT Press, Dec. 1996, pp. 127–133. 1996.

[16] K. Fukumizu and S. Amari, "Local minima and plateaus in hierarchical structures of multilayer perceptrons," *Neural Netw.*, vol. 13, no. 3, pp. 317–327, 2000.

[17] H. Wei and S.-I. Amari, "Dynamics of learning near singularities in radial basis function networks," *Neural Netw.*, vol. 21, no. 7, pp. 989–1005, Sep. 2008.

[18] W. Guo, H. Wei, J. Zhao, and K. Zhang, "Averaged learning equations of error-function-based multilayer perceptrons," *Neural Comput. Appl.*, vol. 25, nos. 3–4, pp. 825–832, Sep. 2014.

[19] F. Cousseau, T. Ozeki, and S.-I. Amari, "Dynamics of learning in multilayer perceptrons near singularities," *IEEE Trans. Neural Netw.*, vol. 19, no. 8, pp. 1313–1328, Aug. 2008.

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

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

[22] H. Park, M. Inoue, and M. Okada, "Online learning dynamics of multilayer perceptrons with unidentifiable parameters," *J. Phys. A, Math. Gen.*, vol. 36, no. 47, p. 11753, 2003.

[23] W. Guo, H. Wei, Y.-S. Ong, J. R. Hervas, J. Zhao, H. Wang, and K. Zhang, "Numerical analysis near singularities in RBF networks," *J. Mach. Learn. Res.*, vol. 19, no. 1, pp. 1–39, 2018.

[24] S. I. Amari, "Natural gradient works efficiently in learning," *Neural Comput.*, vol. 10, no. 2, pp. 251–276, 1998.

[25] H. Park, S.-I. Amari, and K. Fukumizu, "Adaptive natural gradient learning algorithms for various stochastic models," *Neural Netw.*, vol. 13, no. 7, pp. 755–764, Sep. 2000.

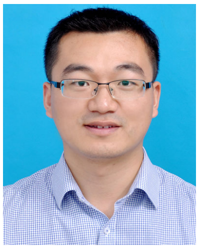
[26] J. Zhao, H. Wei, C. Zhang, W. Li, W. Guo, and K. Zhang, "Natural gradient learning algorithms for RBF networks," *Neural Comput.*, vol. 27, no. 2, pp. 481–505, Feb. 2015.

[27] B. R. Grosse and R. Salakhutdinov, "Scaling up natural gradient by sparsely factorizing the inverse Fisher matrix," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, Lille, France, F. R. Bach and D. M. Blei, Eds. vol. 37, Jul. 2015, pp. 2304–2313.

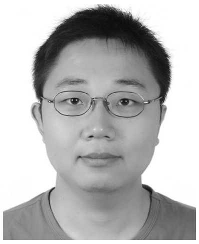
[28] R. Pascanu and Y. Bengio, "Revisiting natural gradient for deep networks," in *Proc. 2nd Int. Conf. Learn. Represent. (ICLR)*, Banff, AB, Canada, Y. Bengio and Y. LeCun, Eds. Apr. 2014, pp. 1–18.

[29] H. Park and K. Lee, "Adaptive natural gradient method for learning of stochastic neural networks in mini-batch mode," *Appl. Sci.*, vol. 9, no. 21, p. 4568, Oct. 2019.

[30] Z. Liao, T. Drummond, I. Reid, and G. Carneiro, "Approximate Fisher information matrix to characterise the training of deep neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 15–26, Jan. 2020.



WEILI GUO received the B.S. degree from the School of Science, Shandong Jianzhu University, China, in 2007, the M.S. degree from the School of Science, Nanjing Agricultural University, China, in 2010, and the Ph.D. degree from the School of Automation, Southeast University, China, in 2014. From 2016 to 2017, he was a Postdoctoral Fellow with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. He is currently working as an Associate Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. His research interests include singular learning dynamics of neural networks, deep learning, and machine learning.



GUANGYU LI received the B.S. degree from the China University of Mining and Technology, in 2008, the M.S. degree from Tongji University, China, in 2011, and the Ph.D. degree from the University of Paris-Sud, Paris, France, in 2015. He is currently working as an Assistant Professor with the Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, Nanjing University of Science and Technology, Nanjing, China. His research interests include information dissemination in vehicle networks, big data mining, electric vehicles charging/discharging scheduling strategy, and traffic control.

JIANFENG LU (Member, IEEE) received the B.S. degree in computer software and the M.S. and Ph.D. degrees in pattern recognition and intelligent system from the Nanjing University of Science and Technology, Nanjing, China, in 1991, 1994, and 2000, respectively. He is currently a Professor and the Vice Dean of the School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interests include image processing, pattern recognition, and data mining.