IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 23, 2018, accepted September 28, 2018, date of publication October 4, 2018, date of current version November 8, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2873811

Influence Area of Overlap Singularity in Multilayer Perceptrons

WEILI GUO[1,2], YUAN YANG[1], YINGJIANG ZHOU[3], YUSHUN TAN[4], HAIKUN WEI[4], AIGUO SONG[1], AND GUOCHEN PANG[2]

[1]Key Laboratory of Remote Measurement and Control of Jiangsu Province, School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China
[2]School of Automation and Electrical Engineering, Linyi University, Linyi 276005, China
[3]School of Automation, Nanjing University of Posts and Telecommunications, Nanjing 210023, China
[4]Key Laboratory of Measurement and Control of CSE, School of Automation, Ministry of Education, Southeast University, Nanjing 210096, China

Corresponding author: Weili Guo (weiliguo@seu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61773118, Grant 61603196, and Grant 61601123, in part by the Natural Science Foundation of Jiangsu Province of China under Grant BK20150851 and Grant BK20160696, in part by the Jiangsu Planned Projects for Postdoctoral Research Funds under Grant 1601001A, in part by the Open Project Program of Jiangsu Key Lab of Remote Measurement and Control under Grant 2242018K30005, and in part by the Fundamental Research Funds for the Central Universities.

ABSTRACT The existing overlap singularities in the parameter space significantly affect the learning dynamics of the multilayer perceptrons. From the obtained theoretical learning trajectories near overlap singularity, when the learning process has been affected by the overlap singularity, the influence area of the overlap singularity is just the line space where the two hidden units equal to each other. However, in the practical applications, different case has been observed and the influence area of such singularity may be larger. By analyzing the generalization error of multilayer perceptrons, we find that the error surface is much flatter near overlap singularity and the singularity would have much larger influence area. Finally, the validity of the obtained results are verified by taking an artificial experiment and two real-data experiments.

INDEX TERMS Multilayer perceptrons, dynamics, information geometry, overlap singularity, influence area.

I. INTRODUCTION

The results in [1] indicate that there exist singular regions in the parameter spaces for almost all learning machines, and the singularities are the subspaces where the Fisher information matrix degenerates [2], [3]. As a typical type of learning machines, feedforward neural networks have been widely used in many fields [4]–[7]. Due to the existence of the singularities, the learning dynamics of neural networks often present strange behaviors. For example, the learning process may become very slow and plateau phenomenon often occurs (an example is shown in Fig. 1) [8]. Also because of the singularities, the standard statistical paradigm of the Cramér-Rao theorem does not hold [9], [10] and the classical model selection criteria, such as Akaike information criterion (AIC), Bayes information criterion (BIC) and minimum description length (MDL), often fail in determining appropriate network structure [11].

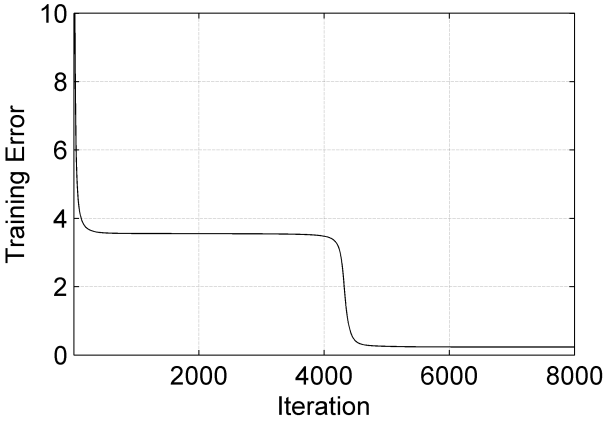


FIGURE 1. Plateau phenomenon occurred in the learning process.

Many researchers have investigated the learning dynamics near singularities of feedforward neural networks, such as multilayer perceptrons(MLPs) [12]–[15], radial basis function (RBF) networks [16], [17], Gaussian mixtures [18], ect, and plenty of results have been obtained. These theoretical analysis results all indicate that the singularities seriously affect the learning dynamics of feedforward neural

2169-3536 © 2018 IEEE. Translations and content mining are permitted for academic research only.
Personal use is also permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

networks and make us further recognize the essence of the singularities.

In recent years, deep learning has become a very hot topic in the machine learning community [19]. Deep neural networks are designed based on traditional neural networks [20]–[22]. Due to the much larger number of hidden layers and architecture size, training deep neural networks also faces many challenges [23], [24]. [25] investigates the deep linear neural networks and finds that the error does not change under a scaling transformation. This would cause the training difficulty which is called scaling symmetries in [26] and [27]. Reference [27], [28] investigate the influence of singularities in deep neural networks. These results all indicate that the singularities also seriously affect the learning dynamics in deep neural networks.

Although many results about the learning dynamics near the singularities have been obtained, the size of the influence area of singularities is still unknown. Reference [29] takes a general mathematical analysis of the learning dynamics near singularities in layered networks and obtains the common learning trajectories near overlap singularities which are represented by a very simple form. The theoretical learning trajectories indicate that when the learning process is affected by the overlap singularity, the student parameters always arrive in such singularity precisely, namely two units finally overlap exactly. However, for MLPs, we find different case in practical simulation experiments, the influence area of the overlap singularity is larger than the theoretical analysis.

In this paper, we aim to investigate the influence area of the overlap singularity in MLPs by analyzing the generalization error surface. The remainder of the paper is organized as follows. In Section 2, we give a more detailed introduction to the motivation of this paper. The generalization error surface of the MLPs near the overlap singularity is analyzed in Section 3. Section 4 is devoted to the simulations and Section 5 states conclusions and discussions.

II. LEARNING PARADIGM

Here, we firstly introduce a typical learning paradigm of MLPs using the standard gradient descent algorithm to minimize the mean square error loss function. For a typical MLP with single hidden layer, it accepts an input vector $\boldsymbol{x}$ and gives a scalar output, i.e.:

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

where k denotes the hidden unit number, $\boldsymbol{J}_i \in \mathcal{R}^n$, $i = 1, \cdots, k$ denotes the weight from the input layer to the ith hidden unit and $w_i \in \mathcal{R}$ denotes the weight from the ith hidden unit to the output layer; n denotes the number of input nodes and $\phi(\boldsymbol{x}, \boldsymbol{J}_i) = \phi(\boldsymbol{J}_i^T \boldsymbol{x})$ denotes an activation function. $\boldsymbol{\theta} = \{\boldsymbol{J}_1, \cdots, \boldsymbol{J}_k, w_1, \cdots, w_k\}$ represents all the parameters of the model (1).

Now we introduce two types of singularities [8], [29]. If two hidden units i and j overlap, i.e. $\boldsymbol{J}_i = \boldsymbol{J}_j$, $w_i\phi(\boldsymbol{x}, \boldsymbol{J}_i) + w_j\phi(\boldsymbol{x}, \boldsymbol{J}_j) = (w_i + w_j)\phi(\boldsymbol{x}, \boldsymbol{J}_i)$ remains the same value when $w_i + w_j$ takes a fixed value, regardless of particular values of w_i and w_j. Therefore, we can identify their sum $w = w_i + w_j$, nevertheless, each of w_i and w_j remains unidentifiable. When $w_i = 0$, $w_i\phi(\boldsymbol{x}, \boldsymbol{J}_i) = 0$, whatever value $\boldsymbol{J}_i$ takes. So there are mainly two types of singular regions in the parameter space of the unipolar activation function based MLPs as follows [30]:
1) Overlap singularity:

$$\mathcal{R}_1 = \{\boldsymbol{\theta} | \boldsymbol{J}_i = \boldsymbol{J}_j\}, \tag{2}$$

2) Elimination singularity:

$$\mathcal{R}_2 = \{\boldsymbol{\theta} | w_i = 0\}. \tag{3}$$

In the case of regression, we have a number of observed data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_t, y_t)$, which are generated by:

$$y = f_0(\boldsymbol{x}) + \varepsilon, \tag{4}$$

where $\boldsymbol{x} \in \mathcal{R}^n$, $y \in \mathcal{R}$, and $f_0(\boldsymbol{x})$ is an unknown true generating function (which is called the teacher function). ε is an additive noise, usually subject to Gaussian distribution with zero mean. $f_0(\boldsymbol{x})$ can be approximated by a MLP, which is the student neural model with the form of model (1).

Since the MLPs have universal approximation ability, we can also assume that the teacher model is described by a MLP with s hidden units:

$$y = f_0(\boldsymbol{x}) + \varepsilon = f_0(\boldsymbol{x}, \boldsymbol{\theta}_0) + \varepsilon = \sum_{i=1}^{s} v_i \phi(\boldsymbol{x}, \boldsymbol{t}_i) + \varepsilon, \tag{5}$$

where $\boldsymbol{t}_i \in \mathcal{R}^n$ and $v_i \in \mathcal{R}$ denote the weight parameters connected to the ith hidden unit, and $\boldsymbol{\theta}_0 = (\boldsymbol{t}_1, \cdots, \boldsymbol{t}_s, v_1, \cdots, v_s)$ is the teacher parameter.

The training input is subject to Gaussian distribution with mean zero and covariance identity matrix $\mathbf{I}_n$:

$$q(\boldsymbol{x}) = (\sqrt{2\pi})^{-n} \exp\left(-\frac{\|\boldsymbol{x}\|^2}{2}\right), \tag{6}$$

and the loss function is defined as:

$$l(y, \boldsymbol{x}, \boldsymbol{\theta}) = \frac{1}{2}(y - f(\boldsymbol{x}, \boldsymbol{\theta}))^2. \tag{7}$$

Then by using the gradient descent method to minimize the above loss, the training process can be completed and the learning trajectories can be obtained.

From the results in [29], the theoretical learning trajectories near $\mathcal{R}_1$ are:

$$h = \frac{2w^*}{3} \log \frac{(z^2 + 3)^2}{|z|} + C, \tag{8}$$

where C is a constant depending on the initial model parameter $(h^{(0)}, z^{(0)})$ and

$$h = \frac{1}{2}\boldsymbol{u}^T \boldsymbol{u}, \tag{9}$$

$$\boldsymbol{u} = \boldsymbol{J}_i - \boldsymbol{J}_j, \tag{10}$$

$$z = \frac{w_i - w_j}{w_i + w_j}. \tag{11}$$

The trajectories are shown in Fig. 2.

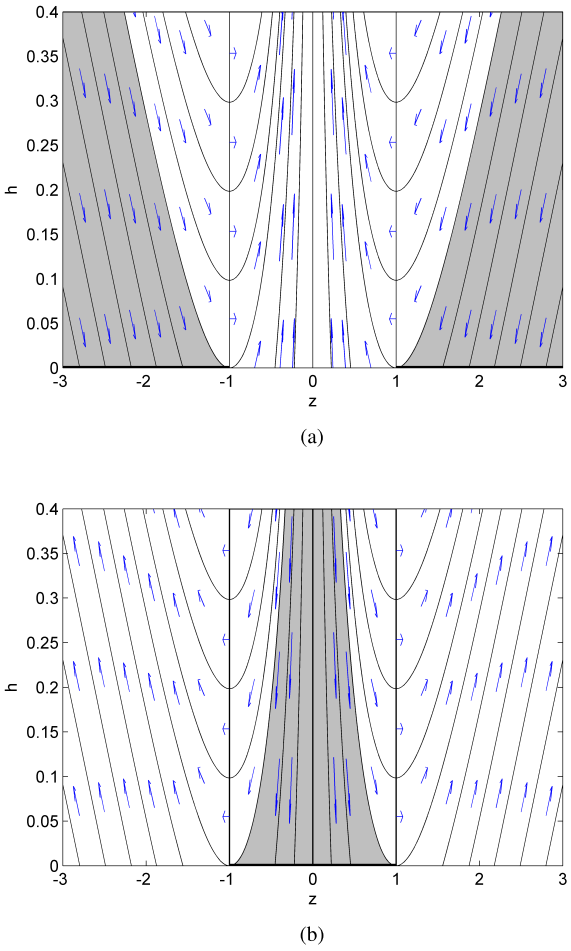


FIGURE 2. Analytical dynamic vector fields. (a) The part $|z| > 1$ is stable. (b) The part $|z| < 1$ is stable.

For the overlap singularity, $\boldsymbol{J}_1 = \boldsymbol{J}_2$, we have $\boldsymbol{u} = 0$, namely $h = 0$. For the elimination singularity, w_1(or w_2) $= 0$, we have $z = -1$(or $+1$). Thus the line $h = 0$ represents the overlap singularity, and the lines $z = \pm 1$ represent the elimination singularity. The overlap singularity is partially stable where the stable area(thick black area in Fig. 2(a) and Fig. 2(b), respectively) is determined by $H(w^*, \boldsymbol{J}^*) = \frac{1}{4} w^* \left\langle e(y, \boldsymbol{x}, \boldsymbol{\theta}) \frac{\partial^2 \phi(\boldsymbol{x}, \boldsymbol{J})}{\partial \boldsymbol{J} \partial \boldsymbol{J}^T} \right\rangle |_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$ [29].

From the trajectory $h \sim z$ shown in Fig. 2 which is obtained by theoretical analysis, for the learning processes which are affected by the overlap singularity, the student parameters always arrive in the line $h = 0$, namely the two hidden units finally overlap exactly. However, the practical situation is different, the influence area might be larger than the theoretical analysis. Next we analyse the generalization error $L(\boldsymbol{\theta})$ at first.

III. THEORETICAL ANALYSIS OF ERROR SURFACE NEAR OVERLAP SINGULARITY

In this section, we analyse the generalization error near the overlap singularity for MLPs and show that the generalization error surface is much flatter near the overlap singularity. Just as pointed out in [16], it is enough to investigate the model with two hidden units for capturing the essence of the learning dynamics near the singularities. Without loss of generality, we analyse the case that both the teacher model and the student model have two hidden units, namely the teacher model and student model have the following form, respectively:

$$f(\boldsymbol{x}, \boldsymbol{\theta}_0) = v_1 \phi(\boldsymbol{x}, \boldsymbol{t}_1) + v_2 \phi(\boldsymbol{x}, \boldsymbol{t}_2), \tag{12}$$

and

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = w_1 \phi(\boldsymbol{x}, \boldsymbol{J}_1) + w_2 \phi(\boldsymbol{x}, \boldsymbol{J}_2). \tag{13}$$

For a given sample set, the corresponding error surface can be obtained by calculating the loss function $l(y, \boldsymbol{x}, \theta)$. Given that the error surface of the loss function cannot avoid the disturbance of the samples, in order to overcome this problem, we can investigate the generalization error $L(\boldsymbol{\theta})$ of MLPs instead:

$$L(\boldsymbol{\theta}) = \langle l(y, \boldsymbol{x}, \boldsymbol{\theta}) \rangle, \tag{14}$$

where $\langle \cdot \rangle$ denotes the average over $(y_t, \boldsymbol{x}_t)$ with respect to the teacher distribution,

$$p_0(y, \boldsymbol{x}) = q(\boldsymbol{x}) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} (y - f_0(\boldsymbol{x}))^2 \right). \tag{15}$$

Since the overlap singularity is basically related with the weights $\boldsymbol{J}_i$, $i = 1, 2$, we can mainly focus on the weights $\boldsymbol{J}_i$, not w_i. Thus in order to quantitatively analyse and visualize the generalization error, without loss of generality, we investigate the case that the student output weights are fixed to the teacher output weights and the dimension of the input is chosen to be 1 in this paper, namely we set $w_1 = v_1$ and $w_2 = v_2$, then the teacher and student model are of the following forms:

$$f(x, \boldsymbol{\theta}_0) = v_1 \phi(x, t_1) + v_2 \phi(x, t_2) + \varepsilon, \tag{16}$$

and

$$f(x, \boldsymbol{\theta}) = v_1 \phi(x, J_1) + v_2 \phi(x, J_2), \tag{17}$$

respectively.

As the output weights have been set to the optimal values, the two parameters v_1 and v_2 do not participate in the learning process and only J_1 and J_2 need to be modified. Thus the system parameters have become $\boldsymbol{\theta} = [J_1, J_2]^T$.

Next, we can quantitatively analyse the generalization error surface of MLPs. When the student parameters arrives in the overlap singularity $\mathcal{R}^* = \{\boldsymbol{\theta}^* | J_1 = J_2 = J^*\}$, then for an arbitrary student parameter $\hat{\boldsymbol{\theta}} = [\hat{J}_1, \hat{J}_2]$, which is near the overlap singularity $\boldsymbol{\theta}^* = [J^*, J^*]$, it can be seen as adding an bias term to $\boldsymbol{\theta}^*$, namely $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + \Delta\boldsymbol{\theta}^*$, where $\Delta\boldsymbol{\theta}^* = \hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = [\hat{J}_1 - J^*, \hat{J}_2 - J^*]^T$.

Then by taking the Taylor expansion of $L(\hat{\boldsymbol{\theta}})$ at $\boldsymbol{\theta}^*$, we have:

$$L(\hat{\boldsymbol{\theta}}) = L(\boldsymbol{\theta}^*) + (\Delta\boldsymbol{\theta}^*)^T \frac{\partial L(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^*} + \frac{1}{2} (\Delta\boldsymbol{\theta}^*)^T \frac{\partial^2 L(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^* \partial \boldsymbol{\theta}^{*T}} \Delta\boldsymbol{\theta}^* + O(\|\Delta\boldsymbol{\theta}^*\|^3). \tag{18}$$

As shown in [30], for the overlap singularity $\boldsymbol{\theta}^*$, we have $\frac{\partial L(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^*} = \mathbf{0}$, then Eq. (18) can be rewritten as:

$$\begin{aligned} L(\hat{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) &= \frac{1}{2}(\Delta\boldsymbol{\theta}^*)^T \frac{\partial^2 L(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^* \partial \boldsymbol{\theta}^{*T}} \Delta\boldsymbol{\theta}^* + O(\|\Delta\boldsymbol{\theta}^*\|^3) \\ &= \frac{1}{2}\Delta\boldsymbol{\theta}^{*T} H(\boldsymbol{\theta}^*)\Delta\boldsymbol{\theta}^* + O(\|\boldsymbol{\theta}^*\|^3), \end{aligned} \tag{19}$$

where $H(\boldsymbol{\theta}) = \dfrac{\partial^2 L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}$ is the Hessian matrix of the MLPs.

For Eq. (19), we have:

$$\begin{aligned} \Delta\boldsymbol{\theta}^{*T} H(\boldsymbol{\theta}^*)\Delta\boldsymbol{\theta}^* &= \Delta\boldsymbol{\theta}^{*T} \begin{bmatrix} \dfrac{\partial^2 L(\boldsymbol{\theta}^*)}{\partial J^{*2}} & \dfrac{\partial^2 L(\boldsymbol{\theta}^*)}{\partial J^{*2}} \\ \dfrac{\partial^2 L(\boldsymbol{\theta}^*)}{\partial J^{*2}} & \dfrac{\partial^2 L(\boldsymbol{\theta}^*)}{\partial J^{*2}} \end{bmatrix} \Delta\boldsymbol{\theta}^* \\ &= \Delta\boldsymbol{\theta}^{*T} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \Delta\boldsymbol{\theta}^* \frac{\partial^2 L(\boldsymbol{\theta}^*)}{\partial J^{*2}}, \end{aligned} \tag{20}$$

Moreover,

$$\begin{aligned} &\Delta\boldsymbol{\theta}^{*T} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \Delta\boldsymbol{\theta}^* \\ &= \begin{bmatrix} \hat{J}_1 - J^*, & \hat{J}_2 - J^* \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \hat{J}_1 - J^*, & \hat{J}_2 - J^* \end{bmatrix}^T \\ &= [(\hat{J}_1 - J^*) + (\hat{J}_2 - J^*), (\hat{J}_1 - J^*) + (\hat{J}_2 - J^*)] \begin{bmatrix} \hat{J}_1 - J^* \\ \hat{J}_2 - J^* \end{bmatrix} \\ &= (\hat{J}_1 - J^*)^2 + 2(\hat{J}_1 - J^*)(\hat{J}_2 - J^*) + (\hat{J}_2 - J^*)^2 \\ &\leq 2((\hat{J}_1 - J^*)^2 + (\hat{J}_2 - J^*)^2) \\ &= 2\|\Delta\boldsymbol{\theta}^*\|^2. \end{aligned} \tag{21}$$

Then we have:

$$L(\hat{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) \leq \frac{\partial^2 L(\boldsymbol{\theta}^*)}{\partial J^{*2}} \|\Delta\boldsymbol{\theta}^*\|^2 + O(\|\boldsymbol{\theta}^*\|^3). \tag{22}$$

Remark 1: From Eq. (14), it is obvious that $L(\theta) \geq 0$. Given that the points in the overlap singularity are all local minima [29], [31], then for an arbitrary point near the overlap singularity, we can obtain $L(\hat{\boldsymbol{\theta}}) > L(\boldsymbol{\theta}^*)$, i.e. $L(\hat{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) > 0$.

By taking some calculations, we can obtain the following results:

Theorem 1: The difference of the generalization error between the point $\hat{\boldsymbol{\theta}}$ around the overlap singularity and the overlap singularity $\boldsymbol{\theta}^*$ satisfies the following in-equation:

$$L(\hat{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) < \frac{2.31}{\pi}(|v_1| + |v_2|)^2 \|\Delta\boldsymbol{\theta}^*\|^2 + O(\|\boldsymbol{\theta}^*\|^3). \tag{23}$$

Proof: The calculation process is shown in Appendix. □

From Theorem 1, we can see that the generalization error near the overlap singularity changes by the order of $O(\|\Delta\boldsymbol{\theta}^*\|^2)$ where the coefficient is small than $\frac{2.31}{\pi}(|v_1| + |v_2|)^2$. The nearer the student parameters to the overlap singularities, the smaller the distance between the two hidden units. Especially when $\|\Delta\boldsymbol{\theta}^*\|^2 < 1$, the difference between $L(\hat{\boldsymbol{\theta}})$ and $L(\boldsymbol{\theta}^*)$ is much less than the distance between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$, which implies that the generalization error near the overlap singularity is much flatter. Thus, when the student parameters arrive in the space near the overlap singularity, the variation of the generalization error is much less which leads to the parameters change slightly. Even though the two hidden units are not precisely equal to each other, the learning process is still affected by the overlap singularity. Thus the influence area of the overlap singularity in MLPs is larger than the theoretical results which is only the subspace $\mathcal{R}^* = \{\boldsymbol{\theta}^* | \boldsymbol{J}_1 = \boldsymbol{J}_2\}$.

To verify the above analysis, we carry out two experiments involving different cases in the simulation part.

IV. SIMULATION PART

In this section, we take three experiments to verify the above analytical results. In Experiment 1, we focus on the artificial case that the teacher model is described by MLPs. Then in Experiments 2 and 3, we consider the factual case that two real datasets is approximated by the MLPs. The simulation results can illustrate the correctness of Theorem 1.

A. GENERALIZATION ERROR SURFACE OF THE MLPs

In this experiment, we consider the case that the teacher model and student model are of the forms in Eq. (16) and Eq. (17), respectively. Since only J_1 and J_2 are the variable parameters, the generalization error surface can be shown visually after the corresponding generalization errors are obtained. In the generalization error surface, the influence of the overlap singularity can be observed directly and clearly.

In order to obtain the generalization error surface, we should get the analytical form of generalization error at first. However, it is hard to deal with this problem because of the non-integrability of the traditional log-sigmoid function $f(x) = \dfrac{1}{1 + \mathrm{e}^{-\lambda x}}$. In order to overcome this problem, we adopt the error funcion $f(x) = \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{-\infty}^{x} \exp(-\frac{t^2}{2})\mathrm{d}t$ as the activation function and use the following averaged learning equation to investigate the learning dynamics of MLPs [31], [32]:

$$\dot{J}_i = -\eta \frac{\partial L(\boldsymbol{\theta})}{\partial J_i}, \tag{24}$$

where $i = 1,\ 2$, and η denotes the learning rate.

By using the obtained results in Eq. (18)-(21) and Eq. (24) [15], we have:

$$\dot{J}_i = \eta v_i \Big(\sum_{j=1}^{2} v_j P_2(t_j, J_i) - \sum_{j=1}^{2} v_j P_2(J_j, J_i)\Big), \quad \text{for } i = 1,\ 2 \tag{25}$$

and the analytical form of generalization error

$$\begin{aligned} L(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{2} v_i v_j P_1(t_i, t_j) - \sum_{i=1}^{2}\sum_{j=1}^{2} v_i v_j P_1(t_i, J_j) \\ + \frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{2} v_i v_j P_1(J_i, J_j), \end{aligned} \tag{26}$$

where:

$$P_1(t, J) = \frac{1}{2\pi} \arcsin \frac{Jt}{\sqrt{1+J^2}\sqrt{1+t^2}} + \frac{1}{4}, \quad (27)$$

and

$$P_2(t, J) = \frac{1}{2\pi} \frac{1}{\sqrt{1+J^2+t^2}} \frac{t}{\sqrt{1+J^2}}. \quad (28)$$

Then for given teacher parameters and initial values of student parameters, the learning process of the student parameters can be obtained by solving Eq. (25).

In this experiment, we choose the teacher parameters $t_1 = 0.48$, $t_2 = -0.85$, $v_1 = 0.60$ and $v_2 = 0.20$. By letting the initial states of the two student parameters be identical, we can obtain the best approximation $J^* = 0.1647$. Then we choose the initial student parameters as $J_1^{(0)} = -0.90$ and $J_2^{(0)} = 0.65$, the final states are $J_1 = 0.1314$ and $J_2 = 0.2684$. The experiment results are shown in Fig. 3, where Fig. 3(a)-(d) represent the surface of $L(\hat{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)$ near the overlap singularity, the trajectory of generalization error, the trajectory of J_1 and J_2, and the generalization error surface near the overlap singularity, respectively. '$\circ$' and '$\times$' represent the initial state and final state, respectively.

As shown in Fig. 3(a), $L(\hat{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)$ is much smaller than $\frac{2.31}{\pi}(|v_1| + |v_2|)^2 \|\Delta\boldsymbol{\theta}^*\|^2$, which verifies the correctness of Theorem 1. From Fig. 3(b), it can be seen that the generalization error decreases fast at the early stage of the learning process, then almost remains unchanged till the end of the training. Meanwhile, J_1 and J_2 become close to each other (shown in Fig. 3(c)). It is clear in Fig. 3(d) that the learning trajectory tends to the overlap singularity. Although the two hidden units do not overlap exactly, the learning dynamics are still influenced by the overlap singularity. After the training process, $\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 1.0e-05 \times [0.2317, -0.7515]^T$, i.e. the gradient becomes very small and the generalization error surface is very flat, thus the student parameters remain slight change even the training process becomes longer, i.e. the influence area of the overlap singularity is much larger.

B. CUFF-LESS BLOOD PRESSURE ESTIMATION

After having taken an artificial experiment to verify the obtained results, next we do two real experiments to verify the validity of the theoretical analysis. In experiment 2, the MLPs are used to approximate the blood pressure estimation database [33]. In this online waveform database, after collecting the photoplethysmograph (PPG) and electrocardiogram (ECG) signal, the arterial blood pressure (ABP) signal can be estimated by using approximation algorithms. Then in this experiment, the input is $\hat{\boldsymbol{x}} = [x_1, x_2]^T$ and the output is y for MLPs, where x_1 is PPG, x_2 is ECG and y is ABP. For machine learning, preprocessing is usually required for obtaining the better performance and we choose the Gaussian normalization in this paper [34]. $\hat{\boldsymbol{x}}(k)$ is normalized as:

$$\boldsymbol{x}(k) = \frac{\hat{\boldsymbol{x}}(k) - \boldsymbol{\mu}}{\boldsymbol{\delta}}, \quad \text{for } k = 1, 2, \cdots, M \quad (29)$$

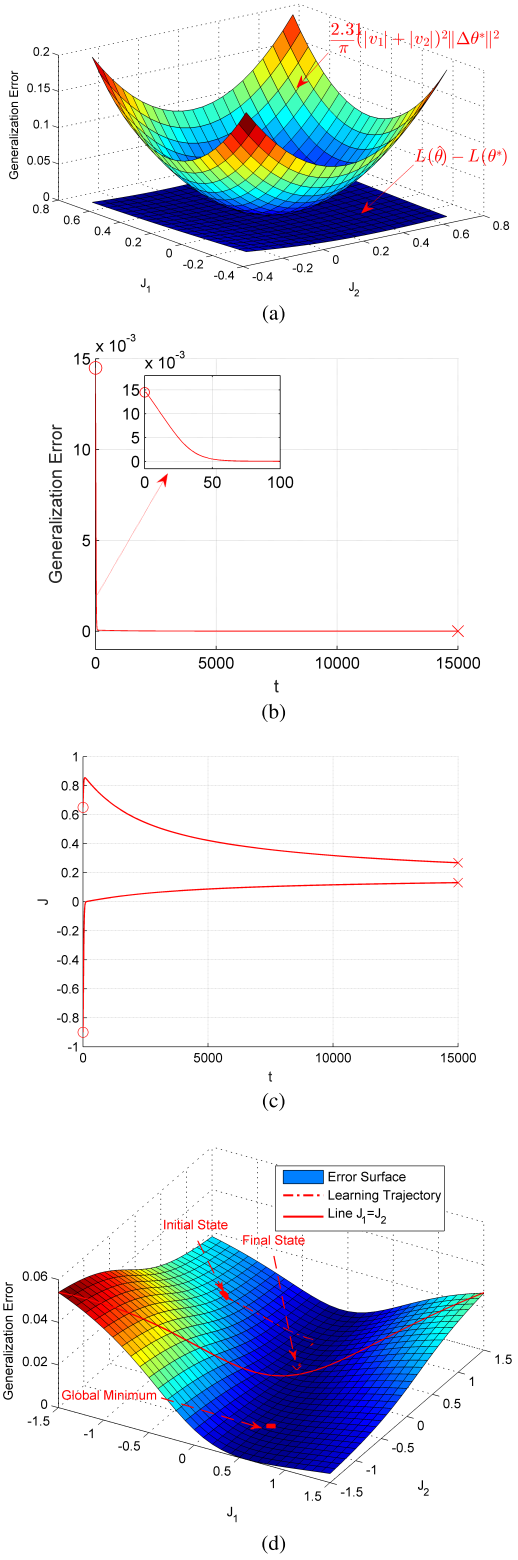


FIGURE 3. Learning trajectories in Experiment 1. (a) The trajectory of J. (b) The generalization error surface. (c) The trajectory of J. (d) The generalization error surface.

where $\boldsymbol{\mu}$ is the sample mean value of the $\boldsymbol{x}$:

$$\boldsymbol{\mu} = \frac{1}{M} \sum_{i=1}^{M} \hat{\boldsymbol{x}}(i), \quad (30)$$

$\boldsymbol{\delta}$ is the sample standard deviation:

$$\boldsymbol{\delta} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(\hat{\boldsymbol{x}}(i) - \boldsymbol{\mu})^2}, \tag{31}$$

and M is the number of sample in the data set.

Different from Experiment 1, as the input distribution is unknown, the ALE is inapplicable. Instead, we use the batch mode learning in the training process. Given that the dimension of input is not 1, in order to directly show how close between two units i and j is, in Experiments 2 and 3 we adopt the squared Euclidean distance $h(i, j)$ which is defined in Eq. (9). The closer the two hidden units are to each other, the closer the squared Euclidean distance between them is to 0. The hidden unit number of the student model is chosen as $k = 8$, namely the student MLP is given by:

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{i=1}^{8} w_i \phi(\boldsymbol{x}, \boldsymbol{J}_i). \tag{32}$$

We use $N = 200$ samples to train the MLPs, and use the sum squared training error to replace the generalization error. Then the model is trained for 10000 times with the learning rate $\eta = 0.03$, and the initial and final states of the student parameters are as follows:

$$\begin{aligned}\boldsymbol{J}^{(0)} &= \left[\boldsymbol{J}_1^{(0)}, \boldsymbol{J}_2^{(0)}, \boldsymbol{J}_3^{(0)}, \cdots, \boldsymbol{J}_8^{(0)}\right] \\ &= \begin{bmatrix} 0.7369 & -0.0981 & -0.0031 & -0.9533 \\ 0.8760 & -0.9857 & 0.9645 & 0.2263 \\ 0.6443 & -0.1546 & -0.9210 & 0.9314 \\ -0.1941 & 0.2266 & -0.6531 & -0.9248 \end{bmatrix},\end{aligned} \tag{33}$$

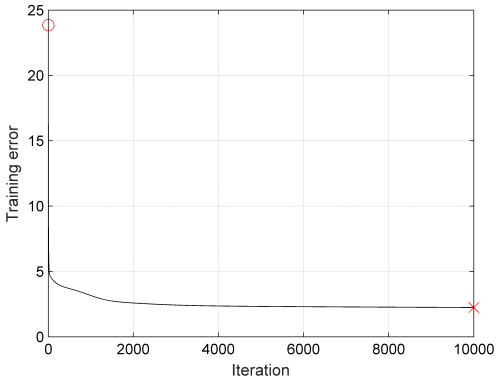
$$\begin{aligned}\boldsymbol{w}^{(0)} &= \left[w_1^{(0)}, w_2^{(0)}, w_3^{(0)}, \cdots, w_8^{(0)}\right] \\ &= [0.3828\ \ 0.9461,\ \ -0.6826,\ \ 0.2903 \\ &\quad -0.7796,\ \ 0.2582,\ \ 0.7194,\ \ -0.3073],\end{aligned} \tag{34}$$

$$\begin{aligned}\boldsymbol{J} &= [\boldsymbol{J}_1, \boldsymbol{J}_2, \boldsymbol{J}_3, \cdots, \boldsymbol{J}_8] \\ &= \begin{bmatrix} 1.2186 & 1.3180 & -1.2851 & -1.1058 \\ 0.9104 & -0.4648 & 0.4644 & 0.3508 \\ 0.4478 & -0.3738 & -2.0233 & 2.8486 \\ -0.7589 & 0.0577 & -1.5594 & -0.8602 \end{bmatrix},\end{aligned} \tag{35}$$

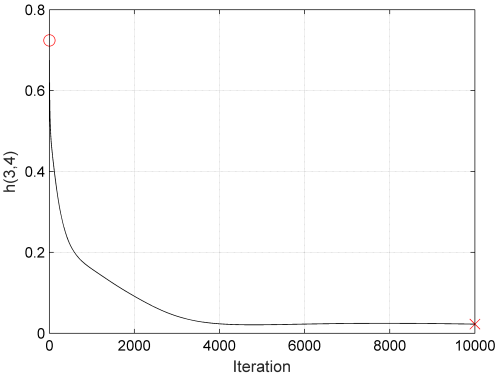
$$\begin{aligned}\boldsymbol{w} &= [w_1, w_2, w_3, \cdots, w_8] \\ &= [0.7465,\ \ 1.5188,\ \ -0.8877,\ \ -0.1907 \\ &\quad -0.5114,\ \ 0.5144,\ \ 0.9126,\ \ -1.8684].\end{aligned} \tag{36}$$

The simulation results are shown in Fig. 4, where Fig. 4(a)-(c) represent the trajectories of training error, $h(3, 4)$ and w_i, $i = 1, \cdots, 8$, respectively. '$\circ$' and '$\times$' represent the initial state and final state, respectively. $h(3, 4)$ is the squared Euclidean distance between hidden nodes 3 and 4, i.e. $h(3, 4) = \frac{1}{2}(\boldsymbol{J}_3 - \boldsymbol{J}_4)^T(\boldsymbol{J}_3 - \boldsymbol{J}_4)$.

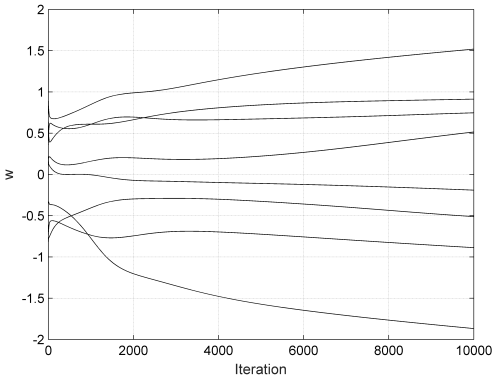
From Fig. 4, we can see that the learning dynamics are similar to the results in Experiment 1. At the early stage of the training process, the training error reduces fast and remains almost unchanged till the end (see Fig. 4(a)). Corresponding to this, $h(3, 4)$ tends to zero rapidly (see Fig. 4)



(a)



(b)



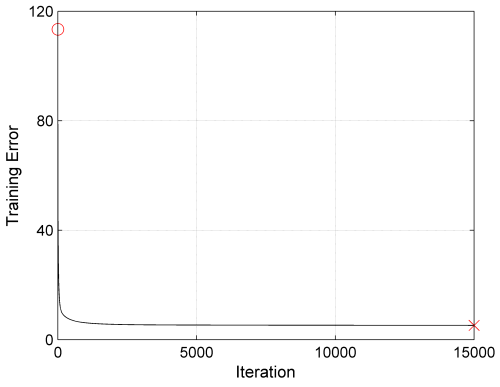
(c)

FIGURE 4. Learning trajectories in Experiment 2. (a) The trajectory of training error. (b) The trajectory of $h(3, 4)$. (c) The trajectory of $\boldsymbol{w}$.

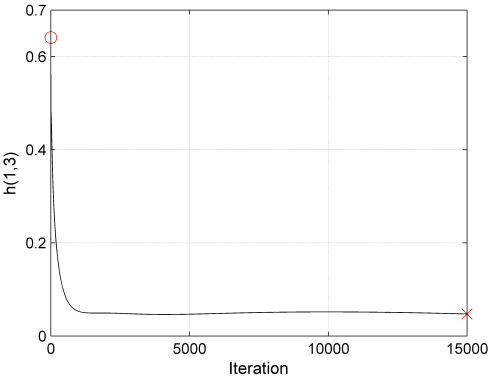
and finally retains a small value (0.0225). From Eq. (35), after training the units $\boldsymbol{J}_3 = [-1.2851,\ 0.4644]^T$ and $\boldsymbol{J}_4 = [-1.1058,\ 0.3508]^T$, $\boldsymbol{J}_3$ and $\boldsymbol{J}_4$ have some difference, however the learning process is still affected by the overlap singularity. This is in accordance to the obtained results that the overlap singularity has larger influence area.

C. COMBINED CYCLE POWER PLANT (CCPP) DATASET

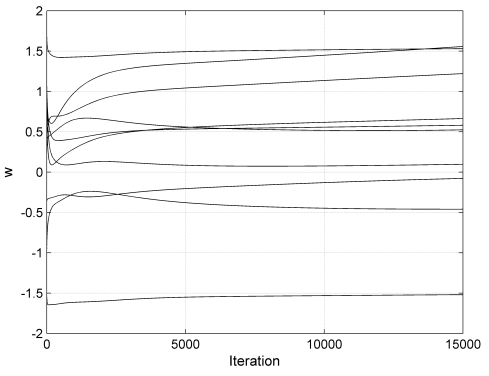
Combined Cycle Power Plant (CCPP) dataset is also a regression benchmark from the UCI Machine Learning Repository [35]–[37]. Hourly average ambient variables



(a)



(b)



(c)

FIGURE 5. Learning trajectories in Experiment 3. (a) The trajectory of training error. (b) The trajectory of $h(1, 3)$. (c) The trajectory of $\boldsymbol{w}$.

Temperature (T), Ambient Pressure (AP), Relative Humidity (RH) and Exhaust Vacuum (V) are used to predict the net hourly electrical energy output (EP) of the plant. Thus for the MLPs, the input is $\boldsymbol{x} = [x_1,\ x_2,\ x_3,\ x_4]^T$, where x_1 is T, x_2 is AP, x_3 is RH and x_4 is V, the output is EP. The preprocessing is also done as the Experiment 2.

The hidden unit number of the student model is chosen as $k = 10$, namely the student MLP is given by:

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{i=1}^{10} w_i \phi(\boldsymbol{x}, \boldsymbol{J}_i). \tag{37}$$

Batch mode learning is used to accomplish the training process and we use $N = 200$ samples to train the MLPs. Then the model is trained for 15000 times with the learning rate $\eta = 0.001$.

Fig. 5 presents the simulation results, where Fig. 5(a)-(c) represent the trajectories of training error, $h(1, 3)$ and w_i, $i = 1, \cdots, 10$, respectively. '$\circ$' and '$\times$' represent the initial state and final state, respectively. $h(1, 3)$ is the squared Euclidean distance between hidden nodes 1 and 2, i.e. $h(1, 3) = \frac{1}{2}(\boldsymbol{J}_1 - \boldsymbol{J}_3)^T(\boldsymbol{J}_1 - \boldsymbol{J}_3)$.

The initial state and final state of hidden unit 1 and 3 are shown as follows:

$$[\boldsymbol{J}_1^{(0)},\ \boldsymbol{J}_3^{(0)}] = \begin{bmatrix} 1.2442 & 0.7505 \\ -0.7139 & 0.1948 \\ 1.9443 & 1.0089 \\ -0.1407 & 0.2051 \end{bmatrix}, \tag{38}$$

$$[w_1^{(0)},\ w_3^{(0)}] = [-1.2276,\ \ 1.7313], \tag{39}$$

and

$$[\boldsymbol{J}_1,\ \boldsymbol{J}_3] = \begin{bmatrix} 0.5509 & 0.4581 \\ -0.2229 & 0.0554 \\ 0.9107 & 0.9899 \\ 0.3372 & 0.2898 \end{bmatrix}, \tag{40}$$

$$[w_1,\ w_3] = [-1.5202,\ \ 0.6636], \tag{41}$$

respectively.

As shown in Fig. 5, the experiment results are similar to those previously obtained in Experiment 2. From the initial values and final values of $\boldsymbol{J}_1$ and $\boldsymbol{J}_3$ (shown in Eq. (38) and Eq. (40)) and the trajectory of $h(1, 3)$ (shown in Fig. 5(b)), although there is some difference between $\boldsymbol{J}_1$ and $\boldsymbol{J}_3$, the learning process is also affected by the overlap singularity.

From the results in Experiments 1 - 3, the correctness of Theorem 1 has been verified. Thus the overlap singularity indeed has much larger influence for MLPs. Researchers should pay more attention to investigate the way to avoid or reduce the influence of the overlap singularity.

V. CONCLUSION

For the widely-used MLPs, there exist overlap singularities in the parameter space, and the overlap singularities seriously affected the learning dynamics of MLPs. In the previous theoretical analysis, under the batch mode learning, the model parameters always arrive and trap in the overlap singularity when affected by such singularity. However, by analyzing the generalization error surface near the overlap singularity, for an arbitrary point $\hat{\boldsymbol{\theta}}$ which is near the overlap singularity and the point $\boldsymbol{\theta}^*$ which is on the overlap singularity, we prove that the difference of generalization error between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$ is attenuated by second-order of $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|$. Thus the generalization error surface is much flatter near the overlap singularity, which leads to that, when the learning process is near the overlap singularity, even though the two hidden units have some difference between each other, the learning dynamics

are still affected by the overlap singularity. Because of the flatness of the generalization error surface near the overlap singularity, the model parameters change slightly even after much longer training. By taking an artificial experiment and two real dataset experiments, the obtained results are verify the validity in the simulation part. Due to its much larger influence, the overlap singularity should be paid more attention to investigate how to avoid or reduce its serious influence in the future.

APPENDIX
PROOF OF THEOREM 1

For simplicity, we introduce the following notations:

$$P_3(t,J) = \left\langle \phi(x,t) \frac{\partial^2 \phi(x,J)}{\partial J^2} \right\rangle, \tag{A-1}$$

$$P_4(t,J) = \left\langle \frac{\partial \phi(x,t)}{\partial t} \frac{\partial \phi(x,J)}{\partial J} \right\rangle. \tag{A-2}$$

From Eq. (14), we can obtain:

$$\begin{aligned}
&\frac{\partial^2 L(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^{*2}} \\
&= (v_1+v_2)^2 \left\langle \frac{\partial \phi(\boldsymbol{x},J^*)}{\partial J^*} \frac{\partial \phi(\boldsymbol{x},J^*)}{\partial J^*} \right\rangle \\
&\quad + (v_1+v_2)^2 \left\langle \phi(\boldsymbol{x},J^*) \frac{\partial^2 \phi(\boldsymbol{x},J^*)}{\partial J^{*2}} \right\rangle \\
&\quad - (v_1+v_2) \left\langle (v_1\phi(\boldsymbol{x},t_1)+v_2\phi(\boldsymbol{x},t_2)) \frac{\partial^2 \phi(\boldsymbol{x},J^*)}{\partial J^{*2}} \right\rangle \\
&= (v_1+v_2)^2 P_4(J^*,J^*) + (v_1+v_2)^2 P_3(J^*,J^*) \\
&\quad - (v_1+v_2)\left(v_1 P_3(t_1,J^*) + v_2 P_3(t_2,J^*)\right) \\
&\le (v_1+v_2)^2 |P_4(J^*,J^*) + P_3(J^*,J^*)| \\
&\quad + |v_1+v_2| \left| v_1 P_3(t_1,J^*) + v_2 P_3(t_2,J^*) \right|.
\end{aligned} \tag{A-3}$$

By using the results of Eq. (60)-(62) in [31], the explicit expressions of $P_3(t,J)$ and $P_4(t,J)$ are given as follows:

For $t \neq J$,

$$\begin{aligned}
P_3(t,J) &= -\frac{1}{2\pi} \frac{Jt}{1+J^2} \frac{1}{\sqrt{1+J^2+t^2}} \\
&\quad \times \frac{2(1+J^2+t^2)+1+J^2}{(1+J^2)(1+J^2+t^2)} \\
&= -\frac{Jt}{2\pi} \frac{3(1+J^2)+2t^2}{(1+J^2)^2(1+J^2+t^2)^{\frac{3}{2}}},
\end{aligned} \tag{A-4}$$

$$\begin{aligned}
P_3(J,J) &= -\frac{1}{2\pi} \frac{1}{1+J^2} \frac{1}{\sqrt{1+2J^2}} \\
&\quad \times \left(1 - \frac{2(2+3J^2)J^2}{(1+J^2)(1+2J^2)} - \frac{1+J^2}{1+2J^2}\right) \\
&= -\frac{J^2}{2\pi} \frac{3+5J^2}{(1+J^2)^2(1+2J^2)^{\frac{3}{2}}},
\end{aligned} \tag{A-5}$$

$$P_4(J,J) = \frac{1}{2\pi} \frac{1}{(1+2J^2)^{\frac{3}{2}}}. \tag{A-6}$$

Then, we have:

$$\begin{aligned}
&P_4(J^*,J^*) + P_3(J^*,J^*) \\
&= \frac{1}{2\pi} \left(\frac{1}{(1+2J^{*2})^{\frac{3}{2}}} - \frac{J^{*2}(3+5J^{*2})}{(1+J^{*2})^2(1+2J^{*2})^{\frac{3}{2}}} \right) \\
&= \frac{1}{2\pi} \frac{(1+J^{*2})^2 - J^{*2}(3+5J^{*2})}{(1+J^{*2})^2(1+2J^{*2})^{\frac{3}{2}}} \\
&= \frac{1}{2\pi} \frac{\frac{17}{16} - (2J^{*2}+\frac{1}{4})^2}{(1+J^{*2})^2(1+2J^{*2})^{\frac{3}{2}}}.
\end{aligned} \tag{A-7}$$

For $0 \le J^{*2} < \frac{\sqrt{17}-1}{8}$, we have $P_4(J^*,J^*)+P_3(J^*,J^*) > 0$. As $P_4(J^*,J^*) > 0$, and $P_3(J^*,J^*) < 0$, then we can obtain:

$$\begin{aligned}
|P_4(J^*,J^*) + P_3(J^*,J^*)| &< P_4(J^*,J^*) \\
&= \frac{1}{2\pi} \frac{1}{(1+2J^{*2})^{\frac{3}{2}}} < \frac{1}{2\pi}.
\end{aligned} \tag{A-8}$$

For $J^{*2} \ge \frac{\sqrt{17}-1}{8}$, we have $P_4(J^*,J^*)+P_3(J^*,J^*) \le 0$. As $P_4(J^*,J^*) > 0$, and $P_3(J^*,J^*) < 0$, then we can obtain:

$$\begin{aligned}
&|P_4(J^*,J^*) + P_3(J^*,J^*)| < |P_3(J^*,J^*)| \\
&= \frac{1}{2\pi} \frac{J^{*2}(3+5J^{*2})}{(1+J^{*2})^2(1+2J^{*2})^{\frac{3}{2}}} < \frac{1}{2\pi} \frac{(1+J^{*2})\cdot 3(1+2J^{*2})}{(1+J^{*2})^2(1+2J^{*2})^{\frac{3}{2}}} \\
&= \frac{1}{2\pi} \frac{3}{(1+J^{*2})(1+2J^{*2})^{\frac{1}{2}}} < \frac{1.62}{2\pi}.
\end{aligned} \tag{A-9}$$

Next we focus on $P_3(t_i,J^*)$, we have:

$$\begin{aligned}
|P_3(t_i,J^*)| &= \frac{1}{2\pi} \frac{|J^*|\cdot|t_i|\cdot 3(1+J^2)+2t^2}{(1+J^2)^2(1+J^2+t^2)^{\frac{3}{2}}} \\
&\le \frac{1}{2\pi} \frac{|J^*|\cdot|t_i|\cdot 3(1+J^2+t^2)}{(1+J^2)^2(1+J^2+t^2)^{\frac{3}{2}}} \\
&= \frac{3}{2\pi} \frac{|J^*||t_i|}{(1+J^2)^2(1+J^2+t^2)^{\frac{1}{2}}} \\
&< \frac{3}{2\pi}.
\end{aligned} \tag{A-10}$$

Then we can get:

$$\begin{aligned}
&|v_1 P_3(t_1,J^*) + v_2 P_3(t_2,J^*)| \\
&\le |v_1||P_3(t_1,J^*)| + |v_2||P_3(t_2,J^*)| \\
&< \frac{3}{2\pi}(|v_1|+|v_2|).
\end{aligned} \tag{A-11}$$

Overall, we have:

For $0 \le J^{*2} < \frac{\sqrt{17}-1}{8}$,

$$\begin{aligned}
\frac{\partial^2 L(\boldsymbol{\theta}^*)}{\partial J^{*2}} &< \frac{(v_1+v_2)^2}{2\pi} + \frac{3}{2\pi}|v_1+v_2|(|v_1|+|v_2|) \\
&< \frac{|v_1|+|v_2|}{2\pi}(|v_1|+|v_2|+3(|v_1|+|v_2|)) \\
&= \frac{2}{\pi}(|v_1|+|v_2|)^2.
\end{aligned} \tag{A-12}$$

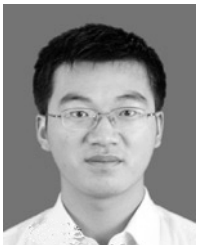
For $J^{*2} \geq \dfrac{\sqrt{17}-1}{8}$,

$$\begin{aligned}\frac{\partial^2 L(\boldsymbol{\theta}^*)}{\partial J^{*2}} &< \frac{1.62}{2\pi}(v_1+v_2)^2 + \frac{3}{2\pi}(|v_1+v_2|(|v_1|+|v_2|) \\ &< \frac{4.62}{2\pi}(|v_1|+|v_2|)^2 \\ &= \frac{2.31}{\pi}(|v_1|+|v_2|)^2. \end{aligned} \tag{A-13}$$

Thus from Eq. (A-12) and Eq. (A-13), $\dfrac{\partial^2 L(\boldsymbol{\theta}^*)}{\partial J^{*2}} < \dfrac{2.31}{\pi}(|v_1|+|v_2|)^2$. This proves Theorem 1. □

REFERENCES

[1] S. Watanabe, "Almost all learning machines are singular," in *Proc. IEEE Symp. Found. Comput. Intell.*, Honolulu, HI, USA, 2007, pp. 383–388.
[2] S. Amari and H. Nagaoka, *Information Geometry*. New York, NY, USA: Oxford Univ. Press, 2000.
[3] S. Amari and T. Ozeki, "Differential and algebraic geometry of multilayer perceptrons," *IEICE Trans. Fund Elect.*, vol. E84, no. 1, pp. 31–38, Jan. 2001.
[4] Y. Guo, "Global asymptotic stability analysis for integro-differential systems modeling neural networks with delays," *Zeitschrift Angewandte Mathematik Physik*, vol. 61, no. 6, pp. 971–978, Dec. 2010.
[5] Y. Guo, "Global stability analysis for a class of Cohen-Grossberg neural network models," *Bull. Korean Math. Soc.*, vol. 49, no. 6, pp. 1193–1198, Nov. 2012.
[6] X. Zhai, A. A. S. Ali, A. Amira, and F. Bensaali, "MLP neural network based gas classification system on Zynq SoC," *IEEE Access*, vol. 4, pp. 8138–8146, 2016.
[7] Y. Guo, "Mean square exponential stability of stochastic delay cellular neural networks," *Electron. J. Qualitative Theory Differ. Equ.*, vol. 2013, no. 34, pp. 1–10, 2013.
[8] 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.
[9] S. Watanabe, "Algebraic analysis for non-identifiable learning machines," *Neural Comput.*, vol. 13, no. 4, pp. 899–933, Apr. 2001.
[10] S. Watanabe, "Algebraic geometrical methods for hierarchical learning machines," *Neural Netw.*, vol. 14, no. 8, pp. 1049–1060, Oct. 2001.
[11] S. Watanabe, "A widely applicable Bayesian information criterion," *J. Mach. Learn. Res.*, vol. 14, pp. 867–897, Mar. 2013.
[12] K. Fukumizu and S. Amari, "Local minima and plateaus in hierarchical structure of multilayer perceptrons," *Neural Netw.*, vol. 13, no. 3, pp. 317–327, Apr. 2000.
[13] 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.
[14] W. Guo, H. Wei, J. Zhao, and K. Zhang, "Theoretical analysis of learning dynamics near the opposite singularities in multilayer perceptrons," (in Chinese), *Control Theory Appl.*, vol. 31, no. 2, pp. 140–147, Feb. 2014.
[15] 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, Aug. 2014.
[16] H. Wei and S. Amari, "Dynamics of learning near singularities in radial basis function networks," *Neural Netw.*, vol. 21, no. 7, pp. 989–1005, Sep. 2008.
[17] 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.
[18] H. Park and T. Ozeki, "Singularity and slow convergence of the EM algorithm for Gaussian mixtures," *Neural Process. Lett.*, vol. 29, no. 1, pp. 45–59, Feb. 2009.
[19] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.
[20] I. Goodfellow, O. Vinyals, and A. M. Saxe. (2014). "Qualitatively characterizing neural network optimization problems." [Online]. Available: http://arxiv.org/abs/1412.6544
[21] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun, "The loss surface of multilayer networks," in *Proc. 18th Int. Conf. Artif. Intell. Statist. (AISTATS)*, San Diego, CA, USA, 2015, pp. 192–204.
[22] Z. C. Lipton. (2016). "Stuck in a what? Adventures in weight space." [Online]. Available: https://arxiv.org/abs/1602.07320
[23] H. van Hasselt, A. Guez, M. Hessel, and D. Silver. (2016). "Learning functions across many orders of magnitudes." [Online]. Available: https://arxiv.org/abs/1602.07714
[24] C. Gulcehre, J. Sotelo, M. Moczulski, and Y. Bengio. (2017). "A robust adaptive stochastic gradient method for deep learning." [Online]. Available: https://arxiv.org/abs/1703.00788
[25] A. Saxe, J. McClelland, and S. Ganguli. (2014). "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks." [Online]. Available: https://arxiv.org/abs/1312.6120
[26] Y. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio, "Identifying and attacking the saddle point problem in high-dimensional non-convex optimization," in *Proc. 27th Adv. Neural Inf. Process. Syst. (NIPS)*, Montreal, QC, Canada, 2014, pp. 2933–2941.
[27] T. Nitta, "On the singularity in deep neural networks," in *Proc. 23rd Int. Conf. Neural Inf. Process.*, Kyoto, Japan, 2016, pp. 389–396.
[28] T. Nitta, "Resolution of singularities introduced by hierarchical structure in deep neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2282–2293, Oct. 2017.
[29] H. Wei, J. Zhang, F. Cousseau, T. Ozeki, and S. Amari, "Dynamics of learning near singularities in layered networks," *Neural Comput.*, vol. 20, no. 3, pp. 813–843, Mar. 2008.
[30] 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.
[31] 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.
[32] D. Saad and A. Solla, "Exact solution for online learning in multilayer neural networks," *Phys. Rev. Lett.*, vol. 74, no. 21, pp. 4337–4340, May 1995.
[33] M. Kachuee, M. Kiani, H. Mohammadzade, and M. Shabany, "Cuff-less high-accuracy calibration-free blood pressure estimation using pulse transit time," in *Proc. IEEE Int. Symp. Circuits Syst.*, Lisbon, Portugal, May 2015, pp. 1006–1009.
[34] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 131–142, Mar. 1998.
[35] P. Tüfekci, "Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods," *Int. J. Electr. Power Energy Syst.*, vol. 60, pp. 126–140, Sep. 2014.
[36] L. Bu, C. Alippi, and D. Zhao, "A pdf-free change detection test based on density difference estimation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 2, pp. 324–334, Feb. 2018.
[37] I. Yamane, H. Sasaki, and M. Sugiyama, "Regularized multitask learning for multidimensional log-density gradient estimation," *Neural Comput.*, vol. 28, no. 7, pp. 1388–1410, Jul. 2016.



WEILI GUO was born in Jining, China, in 1987. He 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 Post-Doctoral Fellow with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. Since 2015, he has been a Post-Doctoral Fellow with the School of Automation, Southeast University, China. His research interests include singular learning dynamics of neural networks, deep learning, and machine learning.

YUAN YANG received the B.S. and M.S. degrees in wireless communication from the China University of Mining and Technology, China, in 2007 and 2010, and the Ph.D. degree from the Institute of Computer Science, Free University of Berlin, Germany, 2015.

Since 2015, she has been a Lecturer with the School of Instrument Science and Engineering, Southeast University, China. Her research interests include non-parametric statistics, parameter estimation, indoor localization, and wireless sensor networks. Her current work also investigates distributed positioning, and target tracking and positioning on indoor mobile robots and drones.

YINGJIANG ZHOU was born in Wuwei, Anhui, China, in 1984. He received the M.S. degree in control theory and control engineering from Hohai University, Nanjing, Jiangsu, in 2010, and the Ph.D. degree in control theory and control engineering from Southeast University, China, in 2014.

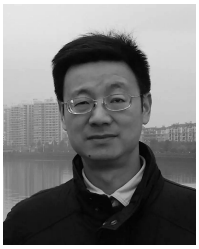
Since 2015, he has been a Lecturer with the School of Automation, Nanjing University of Posts and Telecommunications. His research interests include finite time control of nonlinear systems, network control system, cybersecurity dynamics, and consensus of distributed multi-agent systems and its applications.

YUSHUN TAN was born in Linyi, Shandong, China, in 1977. He received the B.Sc. degree in mathematics and applied mathematics from Qufu Normal University, Qufu, China, in 2003, the M.Sc. degree in probability theory and mathematical statistics from the Huazhong University of Science and Technology, Wuhan, China, in 2006, and the Ph.D. degree in system engineering from Southeast University, Nanjing, in 2015.

Since 2006, he has been a Faculty Member with the College of Applied Mathematics, Nanjing University of Finance and Economics. Since 2016, he has been a Post-Doctoral Researcher with the School of Automation, Southeast University. He is currently a Visiting Fellow with the City University of Hong Kong. His current research interests include networked control systems and complex dynamical networks.

HAIKUN WEI received the B.S. degree from the Department of Automation, North China University of Technology, China, in 1994, and the M.S. and Ph.D. degrees from the Research Institute of Automation, Southeast University, China, in 1997 and 2000, respectively.

From 2005 to 2007, he was a Visiting Scholar with the RIKEN Brain Science Institute, Japan. Since 2000, he has been a Faculty Member with Southeast University, where he is currently a Professor and the Dean of the School of Automation. His research interests include real and artificial in neural networks and industry automation.



AIGUO SONG received the B.S. degree in automatic control and the M.S. degree in measurement and control from Nanjing Aeronautics and Astronautics University, Nanjing, China, in 1990 and 1993, respectively, and the Ph.D. degree in measurement and control from Southeast University, Nanjing, China, in 1996.

From 1996 to 1998, he was a Post-Doctoral Fellow with the Wireless Department, Southeast University. Since 1998, he has been a Faculty Member with Southeast University, where he is currently a Professor and the Director of Robot Sensor and Control Laboratory, and the Dean of the School of Instrument Science and Engineering. His research expertise and interests are in the areas of telerobot, rehabilitation robot, human–computer interface, robot force/tactile sensor, haptic display, and signal processing.

GUOCHEN PANG was born in Linyi, Shandong, China, in 1987. He received the M.S. degree in operational research and cybernetics from Qufu Normal University, Qufu, Shangdong, in 2012, and the Ph.D. degree in control theory and control engineering from Southeast University, China, in 2016. Since 2016, he has been a Lecturer with the School of Automation and Electrical Engineering, Linyi University. His research interests include fault detection, robust control, actuator saturation, and anti-disturbance control and its applications.