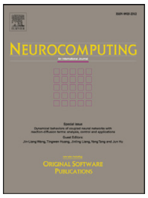
Neurocomputing 282 (2018) 192–201

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Stability analysis of opposite singularity in multilayer perceptrons

Weili Guo[a], Junsheng Zhao[b], Jinxia Zhang[a], Haikun Wei[a,*], Aiguo Song[c], Kanjian Zhang[a]

[a] *Key Laboratory of Measurement and Control of CSE, Ministry of Education, School of Automation, Southeast University, Nanjing 210096, China*
[b] *School of Mathematics Science, Liaocheng University, Liaocheng 252059, China*
[c] *School of Instrument Science & Engineering, Southeast University, Nanjing 210096, China*

ARTICLE INFO

Article history:
Received 12 September 2016
Revised 6 August 2017
Accepted 11 December 2017
Available online 14 December 2017

Communicated by Long Cheng

Keywords:
Stability analysis
Multilayer perceptrons
Opposite singularity
Hessian matrix
Best approximation

ABSTRACT

For the bipolar-activation-function multilayer perceptrons (MLPs), there exist opposite singularities in the parameter space. The Fisher information matrix degenerates on the opposite singularity which causes strange learning behaviors. As the stability is the fundamental to analyze the properties of the opposite singularity, this paper concerns the stability analysis of the opposite singularity in MLPs. The analytical form of the best approximation on the opposite singularity is obtained at first, then the concrete expression of Hessian matrix can be obtained. By analyzing the eigenvalues of Hessian matrix on the opposite singularity, the stability of the opposite singularity is investigated. Finally, two experiments are taken to verify the obtained results.

1. Introduction

As typical learning machines, feedforward neural networks, including multilayer perceptrons (MLPs) and radial basis function (RBF) networks, have been widely used in many fields [1–3]. Different from the learning dynamics in regular statistical models, the learning processes in feedforward neural networks have strange behaviors. For example, the learning behavior in the multilayer perceptrons may be very slow, and plateau phenomenon often occurs in the learning process (an example is shown in Fig. 1) [4–6].

The plateau phenomenon is caused by the existed singular regions in the parameter space [7], and these singularities affect the learning dynamics seriously [8,9]. Meanwhile, the Fisher information matrix of the parameter space degenerates on the singularities [10], which causes the standard gradient descent method no longer Fisher efficient [11]. Then the gradient descent direction is not the steepest descent direction [12,13]. Watanabe [14,15] investigated the effect of singularity in Bayesian inference and proposed a widely applicable Bayesian information criterion (WBIC) which remains efficiency for the singular model [16].

For the multilayer perceptrons, [17] investigated the geometric structure of the parameter space with $H-1$ hidden units embedded in the parameter space of H hidden units and proved that a point in the critical lines corresponding to the global minimum of the smaller model could be a local minimum or a saddle point of the larger model. Guo [18,19] discussed the learning dynamics near overlap singularities of MLPs. By using coordinate transformation and Taylor expansion, [20] gave a general mathematical analysis of the learning dynamics near singularities in layered networks. Based on the methods in [20,21] gave a detailed eigenvalue analysis of the overlap singularity in RBF networks. Park [22] analyzed the dynamics of the EM algorithm for Gaussian mixtures around singularities.

The above obtained results mainly concerned the overlap singularity. For the bipolar-activation-function MLPs, there exist opposite singularities in the parameter space. By using coordinate transformation and Taylor expansion, [23] obtained the analytical form of averaged learning functions, further discussed learning dynamics near the opposite singularity by using simulation experiments.

As is well known, the stability plays a significant role in analyzing the performance of singularities. If the singularity is stable, when the learning process is affected by the singularity, it is hard to escape from it. Thus it is worthy to analyze the stability of the opposite singularity and avoid the stable part of the singularity during the learning process. However, the stability of the opposite singularity has not been investigated yet. In this paper, based on the critical points on the opposite singularity, the stability of the opposite singularity will be analyzed via eigenvalue analysis of Hessian matrix on the opposite singularity.

[*] Corresponding author.
E-mail address: hkwei@seu.edu.cn (H. Wei).

https://doi.org/10.1016/j.neucom.2017.12.018
0925-2312/© 2017 Elsevier B.V. All rights reserved.

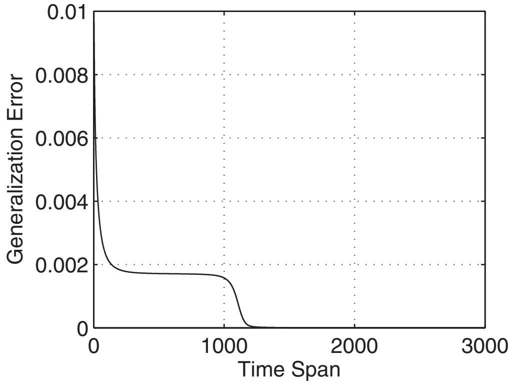


Fig. 1. Plateau phenomenon occurred in the learning process of MLPs.

The rest of this paper is organized as follows. In Section 2, we give a brief introduction of the learning paradigm. The analytical form of the Hessian matrix on the opposite singularity is obtained and the stability of the opposite singularity is analyzed in Section 3. In Section 4, we take two simulation experiments to verify the obtained results. Section 5 states conclusions and discussions.

2. Learning paradigm

Let us consider a typical learning paradigm of MLPs using the standard gradient descent algorithm to minimize the mean square error loss function. 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:

$$y = f_0(\boldsymbol{x}) + \varepsilon, \tag{1}$$

where $\boldsymbol{x} \in \mathcal{R}^N$, $y \in \mathcal{R}$, and $f_0(\boldsymbol{x})$ is an unknown true generating function (called the teacher function). ε is an additive noise that is uncorrelated with training input $\boldsymbol{x}$, usually subjects to Gaussian distribution with zero mean.

The teacher model can be approximated by a student MLP with k hidden units:

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{i=1}^{k} w_i \phi(\boldsymbol{x}, \boldsymbol{J}_i). \tag{2}$$

Here, $\boldsymbol{x} \in \mathcal{R}^N$ denotes the input vector, $\boldsymbol{J}_i \in \mathcal{R}^N$ and $w_i \in \mathcal{R}$ denote the weight parameters connected to the ith hidden unit, N denotes the number of input nodes and $\phi(\boldsymbol{x}, \boldsymbol{J}_i)$ denotes an activation function. $\boldsymbol{\theta} = (\boldsymbol{J}_1, w_1, \ldots, \boldsymbol{J}_k, w_k)$ is the model parameter of the student model.

We assume that the teacher model is also described by a MLP with s hidden units:

$$y = f_0(\boldsymbol{x}) + \varepsilon = f_0(\boldsymbol{x}, \boldsymbol{\theta}_0) + \varepsilon = \sum_{i=1}^{s} v_i \phi(\boldsymbol{x}, \boldsymbol{t}_i) + \varepsilon, \tag{3}$$

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 $\boldsymbol{\theta}_0 = (\boldsymbol{t}_1, v_1, \ldots, \boldsymbol{t}_s, v_s)$ is the teacher parameter. When the true teacher function $f_0(\boldsymbol{x})$ can not be represented by a MLP, $f(\boldsymbol{x}, \boldsymbol{\theta}_0)$ is assumed to be its best approximation by the MLP.

We also assume that the training input is subject to Gaussian distribution with mean zero and covariance identity matrix $\boldsymbol{I}_n$:

$$q(\boldsymbol{x}) = \left(\sqrt{2\pi}\right)^{-n} \exp\left(-\frac{\|\boldsymbol{x}\|^2}{2}\right), \tag{4}$$

and $q(\boldsymbol{x})$ can be generalized to uniform distribution [20].

The loss function is defined as:

$$l(y, \boldsymbol{x}, \boldsymbol{\theta}) = \frac{1}{2}(y - f(\boldsymbol{x}, \boldsymbol{\theta}))^2, \tag{5}$$

and the gradient descent method is used to minimize the above loss.

Since ϕ is an odd function, $(w, \boldsymbol{J})$ and $(-w, -\boldsymbol{J})$ give the same value,

$$w\phi(\boldsymbol{x}, \boldsymbol{J}) = -w\phi(\boldsymbol{x}, -\boldsymbol{J}), \tag{6}$$

namely $w_1\phi(\boldsymbol{x}, \boldsymbol{J}) + w_2\phi(\boldsymbol{x}, -\boldsymbol{J}) = (w_1 - w_2)\phi(\boldsymbol{x}, \boldsymbol{J})$. For this case, we can identify their difference $w = w_1 - w_2$, nevertheless, each of w_1 and w_2 remains unidentifiable.

Thus, there exist opposite singularities in the parameter space as follows:

$$\mathcal{R} = \{\boldsymbol{\theta} | \boldsymbol{J}_i = -\boldsymbol{J}_j\}. \tag{7}$$

For the training methods of MLPs, batch mode learning and online learning are the most commonly used learning modes. In this paper, we use the averaged learning equation (ALE) to investigate the learning dynamics near opposite singularity instead of batch mode learning and online learning. This is because the ALE can be used to investigate not only the batch mode learning dynamics, but also the online learning dynamics [23]. Moreover, by adopting the bipolar error function $\phi(x) = \sqrt{\frac{2}{\pi}}\int_0^x \exp(-\frac{1}{2}t^2)\mathrm{d}t$, namely $\phi(\boldsymbol{x}, \boldsymbol{J}) = \sqrt{\frac{2}{\pi}}\int_0^{\boldsymbol{J}^T\boldsymbol{x}} \exp(-\frac{1}{2}t^2)\mathrm{d}t$, as the activation function of MLPs, the analytical form of the ALE has been obtained in [23]. This will bring us great convenience to take the quantitative analysis of the opposite singularity. Thus we mainly focus on the following ALE:

$$\dot{\boldsymbol{\theta}} = -\eta\left\langle \frac{\partial l(y, \boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\rangle = -\eta \frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \tag{8}$$

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

$$p_0(y, \boldsymbol{x}) = q(\boldsymbol{x}) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y - f_0(\boldsymbol{x}))^2\right), \tag{9}$$

and

$$L(\boldsymbol{\theta}) = \langle l(y, \boldsymbol{x}, \boldsymbol{\theta}) \rangle. \tag{10}$$

The results in [21] indicate that investigating the model with two hidden units is enough to capture the essence of the learning dynamics near the singularities. So we focus on the MLPs with two hidden units, namely the teacher and student models are of the following forms, respectively:

$$f(\boldsymbol{x}, \boldsymbol{\theta}_0) = v_1\phi(\boldsymbol{x}, \boldsymbol{t}_1) + v_2\phi(\boldsymbol{x}, \boldsymbol{t}_2) + \varepsilon, \tag{11}$$

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = w_1\phi(\boldsymbol{x}, \boldsymbol{J}_1) + w_2\phi(\boldsymbol{x}, \boldsymbol{J}_2), \tag{12}$$

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

Guo [23] used the following coordinate transformation to investigate the learning dynamics near the opposite singularities in MLPs:

$$\boldsymbol{v} = \frac{w_2\boldsymbol{J}_2 + w_1\boldsymbol{J}_1}{w_2 - w_1}, \tag{13}$$

$$w = w_2 - w_1, \tag{14}$$

$$\boldsymbol{u} = \boldsymbol{J}_1 + \boldsymbol{J}_2, \tag{15}$$

$$z = \frac{w_2 + w_1}{w_2 - w_1}, \tag{16}$$

and obtained the analytical form of ALEs.

In the new coordinate system, the system parameter becomes $\boldsymbol{\xi} = \{\boldsymbol{v}, w, \boldsymbol{u}, z\}$, and $\mathcal{R}^* = \{\boldsymbol{\xi} | \boldsymbol{u} = \boldsymbol{0}\}$ represents the opposite singularity.

The student model then can be rewritten as:

$$f(\boldsymbol{x}, \boldsymbol{\xi}) = \frac{1}{2}(z-1)w\phi\left(\boldsymbol{x}, \frac{1}{2}(1+z)\boldsymbol{u} - \boldsymbol{v}\right) + \frac{1}{2}(z+1)w\phi\left(\boldsymbol{x}, \frac{1}{2}(1-z)\boldsymbol{u} + \boldsymbol{v}\right). \tag{17}$$

By taking Taylor expansion at $\boldsymbol{u} = \boldsymbol{0}$, we have:

$$f(\boldsymbol{x}, \boldsymbol{\xi}) = w\phi(\boldsymbol{x}, \boldsymbol{v}) + \frac{1}{8}w\left(1-z^2\right)\boldsymbol{u}^T \frac{\partial^2 \phi(\boldsymbol{x}, \boldsymbol{v})}{\partial \boldsymbol{v} \partial \boldsymbol{v}^T}\boldsymbol{u} + O\left(\boldsymbol{u}^3\right). \tag{18}$$

Then Eq. (8) can be rewritten as:

$$\dot{\boldsymbol{\xi}} = -\eta \boldsymbol{T}\boldsymbol{T}^T \frac{\partial L(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}}, \tag{19}$$

where

$$\boldsymbol{T} = \frac{\partial \boldsymbol{\xi}}{\partial \boldsymbol{\theta}^T} = \begin{bmatrix} \frac{z-1}{2} & 0 & 1 & 0 \\ \frac{z+1}{2w}\boldsymbol{u} & -1 & 0 & \frac{z+1}{w} \\ \frac{z+1}{2} & 0 & 1 & 0 \\ \frac{1-z}{2w}\boldsymbol{u} & 1 & 0 & \frac{1-z}{w} \end{bmatrix} \tag{20}$$

is the Jacobi matrix.

Based on the above results, we can do the stability analysis of the opposite singularity $\mathcal{R}^*$ by analyzing the eigenvalues of the Hessian matrix of $L(\boldsymbol{\xi})$.

3. Stability analysis of the opposite singularity

The stability of the opposite singularity can be obtained by evaluating the definiteness of the Hessian matrix on the opposite singularity. In order to solve this problem, we should first obtain the critical points of the opposite singularity and then give the analytical expression of the Hessian matrix on the opposite singularity. Finally, the definiteness of Hessian matrix can be evaluated by analyzing its eigenvalues.

3.1. Critical points on the opposite singularity

In order to obtain the eigenvalues of the Hessian matrix on the opposite singularity, we should obtain the best approximation ($\boldsymbol{v}^*$, w^*) on the opposite singularity at first. When the learning process arrives in the opposite singularity, it can be deemed that the teacher model is approximated by the student model with one unit:

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = w\phi(\boldsymbol{x}, \boldsymbol{v}). \tag{21}$$

The best approximation lies in the opposite singularity which is realized by any point of the following sets:

$$\mathcal{R}^* = \{\boldsymbol{\xi} | \boldsymbol{v} = \boldsymbol{v}^*, w = w^*, \boldsymbol{u} = 0, z \in \mathbb{R}\}. \tag{22}$$

For the best approximation on the opposite singularity, $\boldsymbol{\xi}^* = (\boldsymbol{v}^*, w^*, \boldsymbol{0}, z)$, it satisfies $\frac{\partial L(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}}|_{\boldsymbol{\xi}=\boldsymbol{\xi}^*} = \boldsymbol{0}$. Since $\boldsymbol{u} = \boldsymbol{0}$ is fixed and z can take arbitrary value on the opposite singularity, it is obvious that $\frac{\partial L(\boldsymbol{\xi})}{\partial \boldsymbol{u}}|_{\boldsymbol{\xi}=\boldsymbol{\xi}^*} = \boldsymbol{0}$ and $\frac{\partial L(\boldsymbol{\xi})}{\partial z}|_{\boldsymbol{\xi}=\boldsymbol{\xi}^*} = 0$. Thus the best approximation can be obtained by solving the following equations:

$$\left.\frac{\partial L(\boldsymbol{\xi})}{\partial \boldsymbol{v}}\right|_{\boldsymbol{\xi}=\boldsymbol{\xi}^*} = \boldsymbol{0}, \tag{23}$$

$$\left.\frac{\partial L(\boldsymbol{\xi})}{\partial w}\right|_{\boldsymbol{\xi}=\boldsymbol{\xi}^*} = 0. \tag{24}$$

Now we give the explicit expressions of $\frac{\partial L(\boldsymbol{\xi})}{\partial \boldsymbol{v}}$ and $\frac{\partial L(\boldsymbol{\xi})}{\partial w}$, respectively. For Eqs. (5) and (18), by taking some calculations, we have:

$$\frac{\partial L(\boldsymbol{\xi})}{\partial \boldsymbol{v}} = -w\langle e(y, \boldsymbol{x}, \boldsymbol{\xi}) \frac{\partial \phi(\boldsymbol{x}, \boldsymbol{v})}{\partial \boldsymbol{v}}\rangle + O(\boldsymbol{u}^2), \tag{25}$$

$$\frac{\partial L(\boldsymbol{\xi})}{\partial w} = -\langle e(y, \boldsymbol{x}, \boldsymbol{\xi})\phi(\boldsymbol{x}, \boldsymbol{v})\rangle - \frac{1}{8}(1-z^2) \times \langle e(y, \boldsymbol{x}, \boldsymbol{\xi})\boldsymbol{u}^T \frac{\partial^2 \phi(\boldsymbol{x}, \boldsymbol{v})}{\partial \boldsymbol{v} \partial \boldsymbol{v}^T}\boldsymbol{u}\rangle + O(\boldsymbol{u}^3), \tag{26}$$

where $e(y, \boldsymbol{x}, \boldsymbol{\xi}) = f_0(\boldsymbol{x}) - f(\boldsymbol{x}, \boldsymbol{\xi})$ is the error.

Denote $P_1(\boldsymbol{t}, \boldsymbol{v}) = \langle \phi(\boldsymbol{x}, \boldsymbol{t})\phi(\boldsymbol{x}, \boldsymbol{v})\rangle$ and $P_2(\boldsymbol{t}, \boldsymbol{v}) = \langle \phi(\boldsymbol{x}, \boldsymbol{t})\frac{\partial \phi(\boldsymbol{x}, \boldsymbol{v})}{\partial \boldsymbol{v}}\rangle$, then Eqs. (23) and (24) become:

$$\left.\frac{\partial L(\boldsymbol{\xi})}{\partial \boldsymbol{v}}\right|_{\boldsymbol{\xi}=\boldsymbol{\xi}^*} = w^*(-v_1 P_2(\boldsymbol{t}_1, \boldsymbol{v}^*) - v_2 P_2(\boldsymbol{t}_2, \boldsymbol{v}^*) + w^* P_2(\boldsymbol{v}^*, \boldsymbol{v}^*)) = \boldsymbol{0}, \tag{27}$$

$$\left.\frac{\partial L(\boldsymbol{\xi})}{\partial w}\right|_{\boldsymbol{\xi}=\boldsymbol{\xi}^*} = -v_1 P_1(\boldsymbol{t}_1, \boldsymbol{v}^*) - v_2 P_1(\boldsymbol{t}_2, \boldsymbol{v}^*) + w^* P_1(\boldsymbol{v}^*, \boldsymbol{v}^*) = 0. \tag{28}$$

The explicit expressions of $P_1(\boldsymbol{t}, \boldsymbol{v})$ and $P_2(\boldsymbol{t}, \boldsymbol{v})$ have been obtained in [23]:

$$P_1(\boldsymbol{t}, \boldsymbol{v}) = \frac{2}{\pi} \arcsin \frac{\boldsymbol{t}^T \boldsymbol{v}}{\sqrt{1 + \|\boldsymbol{t}\|^2}\sqrt{1 + \|\boldsymbol{v}\|^2}}, \tag{29}$$

$$P_2(\boldsymbol{t}, \boldsymbol{v}) = \frac{2}{\pi}\sqrt{\det(\boldsymbol{B}(\boldsymbol{t}, \boldsymbol{v})^{-1})}\boldsymbol{A}^{-1}\boldsymbol{t}, \tag{30}$$

where:

$$\boldsymbol{A} = \boldsymbol{I}_n + \boldsymbol{v}\boldsymbol{v}^T, \tag{31}$$

$$\boldsymbol{B}(\boldsymbol{t}, \boldsymbol{v}) = \boldsymbol{A} + \boldsymbol{t}\boldsymbol{t}^T = \boldsymbol{I}_n + \boldsymbol{v}\boldsymbol{v}^T + \boldsymbol{t}\boldsymbol{t}^T, \tag{32}$$

$$\boldsymbol{A}^{-1} = \left(\boldsymbol{I}_n + \boldsymbol{v}\boldsymbol{v}^T\right)^{-1} = \boldsymbol{I}_n - \frac{\boldsymbol{v}\boldsymbol{v}^T}{1 + \|\boldsymbol{v}\|^2}, \tag{33}$$

$$\boldsymbol{B}(\boldsymbol{t}, \boldsymbol{v})^{-1} = \left(\boldsymbol{I}_n - \frac{\boldsymbol{A}^{-1}\boldsymbol{t}\boldsymbol{t}^T}{1 + \boldsymbol{t}^T\boldsymbol{A}^{-1}\boldsymbol{t}}\right)\boldsymbol{A}^{-1}, \tag{34}$$

$$\det\left(\boldsymbol{B}(\boldsymbol{t}, \boldsymbol{v})^{-1}\right) = \frac{1}{(1 + \|\boldsymbol{t}\|^2)(1 + \|\boldsymbol{v}\|^2) - (\boldsymbol{t}^T\boldsymbol{v})^2}, \tag{35}$$

$\boldsymbol{I}_n$ is the compatible identity matrix. (36)

By eliminating the variable w^*, we have:

$$v_1\left(\sqrt{\det(\boldsymbol{B}(\boldsymbol{v}^*, \boldsymbol{v}^*)^{-1})}\,\boldsymbol{v}^* \arcsin \frac{\boldsymbol{t}_1^T \boldsymbol{v}^*}{\sqrt{1 + \|\boldsymbol{t}_1\|^2}\sqrt{1 + \|\boldsymbol{v}^*\|^2}} - \sqrt{\det(\boldsymbol{B}(\boldsymbol{t}_1, \boldsymbol{v}^*)^{-1})}\,\boldsymbol{t}_1 \arcsin \frac{\|\boldsymbol{v}^*\|^2}{1 + \|\boldsymbol{v}^*\|^2}\right)$$
$$-v_2\left(\sqrt{\det(\boldsymbol{B}(\boldsymbol{v}^*, \boldsymbol{v}^*)^{-1})}\,\boldsymbol{v}^* \arcsin \frac{\boldsymbol{t}_2^T \boldsymbol{v}^*}{\sqrt{1 + \|\boldsymbol{t}_2\|^2}\sqrt{1 + \|\boldsymbol{v}^*\|^2}} - \sqrt{\det(\boldsymbol{B}(\boldsymbol{t}_2, \boldsymbol{v}^*)^{-1})}\,\boldsymbol{t}_2 \arcsin \frac{\|\boldsymbol{v}^*\|^2}{1 + \|\boldsymbol{v}^*\|^2}\right) = 0. \tag{37}$$

By solving Eq. (37), we can obtain the best approximation $\boldsymbol{v}^*$, then by substituting $\boldsymbol{v}^*$ to Eq. (27) or (28), the critical points on the opposite singularity $(\boldsymbol{v}^*, w^*)$ can be obtained. Though the form of Eq. (37) is very complex and it is hard to get its analytical solution, the best approximation $(\boldsymbol{v}^*, w^*)$ can still be obtained explicitly when the values of the teacher parameters are fixed.

3.2. Eigenvalue analysis of Hessian matrix on the opposite singularity

Since the best approximation on the opposite singularity has been obtained, the stability of the opposite singularity can be analyzed by taking the eigenvalue analysis of Hessian matrix of MLPs. The Hessian matrix of MLPs based on the averaged loss function is:

$$H(\boldsymbol{\xi}) = \frac{\partial^2 L(\boldsymbol{\xi})}{\partial\boldsymbol{\xi}\partial\boldsymbol{\xi}^T} = \begin{bmatrix} \frac{\partial^2 L(\boldsymbol{\xi})}{\partial\boldsymbol{v}\partial\boldsymbol{v}^T} & \frac{\partial^2 L(\boldsymbol{\xi})}{\partial\boldsymbol{v}\partial w} & \frac{\partial^2 L(\boldsymbol{\xi})}{\partial\boldsymbol{v}\partial\boldsymbol{u}^T} & \frac{\partial^2 L(\boldsymbol{\xi})}{\partial\boldsymbol{v}\partial z} \\ \frac{\partial^2 L(\boldsymbol{\xi})}{\partial w\partial\boldsymbol{v}^T} & \frac{\partial^2 L(\boldsymbol{\xi})}{\partial w^2} & \frac{\partial^2 L(\boldsymbol{\xi})}{\partial w\partial\boldsymbol{u}^T} & \frac{\partial^2 L(\boldsymbol{\xi})}{\partial w\partial z} \\ \frac{\partial^2 L(\boldsymbol{\xi})}{\partial\boldsymbol{u}\partial\boldsymbol{v}^T} & \frac{\partial^2 L(\boldsymbol{\xi})}{\partial\boldsymbol{u}\partial w} & \frac{\partial^2 L(\boldsymbol{\xi})}{\partial\boldsymbol{u}\partial\boldsymbol{u}^T} & \frac{\partial^2 L(\boldsymbol{\xi})}{\partial\boldsymbol{u}\partial z} \\ \frac{\partial^2 L(\boldsymbol{\xi})}{\partial z\partial\boldsymbol{v}^T} & \frac{\partial^2 L(\boldsymbol{\xi})}{\partial z\partial w} & \frac{\partial^2 L(\boldsymbol{\xi})}{\partial z\partial\boldsymbol{u}^T} & \frac{\partial^2 L(\boldsymbol{\xi})}{\partial z^2} \end{bmatrix}. \tag{38}$$

Then we give the explicit expression of $H(\boldsymbol{\xi})$.

For convenience sake, we denote:

$$H(\boldsymbol{\xi}) = \begin{bmatrix} H_{11} & H_{12} & H_{13} & H_{14} \\ H_{21} & H_{22} & H_{23} & H_{24} \\ H_{31} & H_{32} & H_{33} & H_{34} \\ H_{41} & H_{42} & H_{43} & H_{44} \end{bmatrix}, \tag{39}$$

$$P_3(\boldsymbol{t}, \boldsymbol{v}) = \left\langle \phi(\boldsymbol{x}, \boldsymbol{t}) \frac{\partial^2\phi(\boldsymbol{x}, \boldsymbol{v})}{\partial\boldsymbol{v}\partial\boldsymbol{v}^T} \right\rangle, \tag{40}$$

$$P_4(\boldsymbol{v}) = \left\langle \frac{\partial\phi(\boldsymbol{x}, \boldsymbol{v})}{\partial\boldsymbol{v}} \frac{\partial\phi(\boldsymbol{x}, \boldsymbol{v})}{\partial\boldsymbol{v}^T} \right\rangle, \tag{41}$$

and

$$P_5(\boldsymbol{v}) = \left\langle e(y, \boldsymbol{x}, \boldsymbol{\xi}) \frac{\partial^2\phi(\boldsymbol{x}, \boldsymbol{v})}{\partial\boldsymbol{v}\partial\boldsymbol{v}^T} \right\rangle. \tag{42}$$

Based on calculations, we have:

$$\begin{aligned} H_{11} = \frac{\partial^2 L(\boldsymbol{\xi})}{\partial\boldsymbol{v}\partial\boldsymbol{v}^T} &= w^2\left\langle \frac{\partial\phi(\boldsymbol{x}, \boldsymbol{v})}{\partial\boldsymbol{v}} \frac{\partial\phi(\boldsymbol{x}, \boldsymbol{v})}{\partial\boldsymbol{v}^T} \right\rangle \\ &\quad - w\left\langle e(y, x, \boldsymbol{\xi}) \frac{\partial^2\phi(\boldsymbol{x}, \boldsymbol{v})}{\partial\boldsymbol{v}\partial\boldsymbol{v}^T} \right\rangle + O(\boldsymbol{u}^2) \\ &= w^2 P_4(\boldsymbol{v}) - wP_5(\boldsymbol{v}) + O\big(\boldsymbol{u}^2\big), \end{aligned} \tag{43}$$

$$\begin{aligned} H_{12} = \frac{\partial^2 L(\boldsymbol{\xi})}{\partial\boldsymbol{v}\partial w} &= -\left\langle e(y, x, \boldsymbol{\xi}) \frac{\partial\phi(\boldsymbol{x}, \boldsymbol{v})}{\partial\boldsymbol{v}} \right\rangle \\ &\quad + w\left\langle \phi(\boldsymbol{x}, \boldsymbol{v}) \frac{\partial\phi(\boldsymbol{x}, \boldsymbol{v})}{\partial\boldsymbol{v}} \right\rangle + O(\boldsymbol{u}^2) \\ &= -v_1 P_1(\boldsymbol{t}_1, \boldsymbol{v}) - v_2 P_1(\boldsymbol{t}_2, \boldsymbol{v}) + wP_1(\boldsymbol{v}, \boldsymbol{v}) + wP_2(\boldsymbol{v}, \boldsymbol{v}) + O\big(\boldsymbol{u}^2\big), \end{aligned} \tag{44}$$

$$H_{13} = \frac{\partial^2 L(\boldsymbol{\xi})}{\partial\boldsymbol{v}\partial\boldsymbol{u}^T} = \frac{w^2}{4}(1 - z^2)\left\langle \frac{\partial\phi(\boldsymbol{x}, \boldsymbol{v})}{\partial\boldsymbol{v}} \boldsymbol{u}^T \frac{\partial^2\phi(\boldsymbol{x}, \boldsymbol{v})}{\partial\boldsymbol{v}\partial\boldsymbol{v}^T} \right\rangle + O(\boldsymbol{u}^2), \tag{45}$$

$$H_{14} = \frac{\partial^2 L(\boldsymbol{\xi})}{\partial\boldsymbol{v}\partial z} = -\frac{wz}{4}\left\langle \frac{\partial\phi(\boldsymbol{x}, \boldsymbol{v})}{\partial\boldsymbol{v}} \boldsymbol{u}^T \frac{\partial^2\phi(\boldsymbol{x}, \boldsymbol{v})}{\partial\boldsymbol{v}\partial\boldsymbol{v}^T} \boldsymbol{u} \right\rangle + O(\boldsymbol{u}^3), \tag{46}$$

$$H_{21} = H_{12}^T, \tag{47}$$

$$H_{22} = \frac{\partial^2 L(\boldsymbol{\xi})}{\partial w^2} = \langle \phi(\boldsymbol{x}, \boldsymbol{v})\phi(\boldsymbol{x}, \boldsymbol{v}) \rangle = P_1(\boldsymbol{v}, \boldsymbol{v}) + O(\boldsymbol{u}^2), \tag{48}$$

$$\begin{aligned} H_{23} &= \frac{\partial^2 L(\boldsymbol{\xi})}{\partial w\partial\boldsymbol{u}^T} \\ &= \frac{w}{4}(1 - z^2)\boldsymbol{u}^T\left\langle \phi(\boldsymbol{x}, \boldsymbol{v}) \frac{\partial^2\phi(\boldsymbol{x}, \boldsymbol{v})}{\partial\boldsymbol{v}\partial\boldsymbol{v}^T} \right\rangle \\ &\quad - \frac{1}{4}(1 - z^2)\boldsymbol{u}^T\left\langle e(y, \boldsymbol{x}, \boldsymbol{\xi}) \frac{\partial^2\phi(\boldsymbol{x}, \boldsymbol{v})}{\partial\boldsymbol{v}\partial\boldsymbol{v}^T} \right\rangle + O(\boldsymbol{u}^2) \\ &= \frac{1}{4}w(1 - z^2)\boldsymbol{u}^T P_3(\boldsymbol{v}, \boldsymbol{v}) - \frac{1}{4}(1 - z^2)\boldsymbol{u}^T P_5(\boldsymbol{v}) + O(\boldsymbol{u}^2), \end{aligned} \tag{49}$$

$$\begin{aligned} H_{24} = \frac{\partial^2 L(\boldsymbol{\xi})}{\partial w\partial z} &= -\frac{wz}{4}\boldsymbol{u}^T\left\langle \phi(\boldsymbol{x}, \boldsymbol{v}) \frac{\partial^2\phi(\boldsymbol{x}, \boldsymbol{v})}{\partial\boldsymbol{v}\partial\boldsymbol{v}^T} \right\rangle \boldsymbol{u} \\ &\quad + \frac{z}{4}\boldsymbol{u}^T\left\langle e(y, \boldsymbol{x}, \boldsymbol{\xi}) \frac{\partial^2\phi(\boldsymbol{x}, \boldsymbol{v})}{\partial\boldsymbol{v}\partial\boldsymbol{v}^T} \right\rangle \boldsymbol{u} + O(\boldsymbol{u}^3) \\ &= -\frac{wz}{4}\boldsymbol{u}^T P_3(\boldsymbol{v}, \boldsymbol{v})\boldsymbol{u} + \frac{z}{4}\boldsymbol{u}^T P_5(\boldsymbol{v})\boldsymbol{u} + O(\boldsymbol{u}^3), \end{aligned} \tag{50}$$

$$H_{31} = H_{13}^T, \quad H_{32} = H_{23}^T, \tag{51}$$

$$\begin{aligned} H_{33} = \frac{\partial^2 L(\boldsymbol{\xi})}{\partial\boldsymbol{u}\partial\boldsymbol{u}^T} &= \frac{1}{4}w(z^2 - 1)\left\langle e(y, \boldsymbol{x}, \boldsymbol{\xi}) \frac{\partial^2\phi(\boldsymbol{x}, \boldsymbol{v})}{\partial\boldsymbol{v}\partial\boldsymbol{v}^T} \right\rangle + O(\boldsymbol{u}) \\ &= \frac{1}{4}w(z^2 - 1)P_5(\boldsymbol{v}) + O(\boldsymbol{u}), \end{aligned} \tag{52}$$

$$\begin{aligned} H_{34} = \frac{\partial^2 L(\boldsymbol{\xi})}{\partial\boldsymbol{u}\partial z} &= \frac{wz}{2}\left\langle e(y, \boldsymbol{x}, \boldsymbol{\xi}) \frac{\partial^2\phi(\boldsymbol{x}, \boldsymbol{v})}{\partial\boldsymbol{v}\partial\boldsymbol{v}^T} \right\rangle \boldsymbol{u} + O(\boldsymbol{u}^2) \\ &= \frac{1}{2}wzP_5(\boldsymbol{v})\boldsymbol{u} + O(\boldsymbol{u}^2), \end{aligned} \tag{53}$$

$$H_{41} = H_{14}^T, \quad H_{42} = H_{24}^T, \quad H_{43} = H_{34}^T, \tag{54}$$

$$\begin{aligned} H_{44} = \frac{\partial^2 L(\boldsymbol{\xi})}{\partial z^2} &= \frac{1}{4}w\boldsymbol{u}^T\left\langle e(y, \boldsymbol{x}, \boldsymbol{\xi}) \frac{\partial^2\phi(\boldsymbol{x}, \boldsymbol{v})}{\partial\boldsymbol{v}\partial\boldsymbol{v}^T} \right\rangle \boldsymbol{u} + O(\boldsymbol{u}^2) \\ &= \frac{1}{4}w\boldsymbol{u}^T P_5(\boldsymbol{v})\boldsymbol{u} + O(\boldsymbol{u}^3). \end{aligned} \tag{55}$$

Now the analytical form of $H(\boldsymbol{\xi})$ has been obtained. Then after calculating the analytical form of Hessian matrix on the opposite singularity, we can analyze the eigenvalues of $H(\boldsymbol{\xi})$ on the opposite singularity. The results are shown in the following theorem.

Theorem 1. *The stability of the opposite singularity $\mathcal{R}^*$ is determined by $(z^2 - 1)\hat{H}(\boldsymbol{v}^*, w^*)$, where $\hat{H}(\boldsymbol{v}^*, w^*) = \frac{1}{4}w^* P_5(\boldsymbol{v}^*)$. There are three cases:*

⟨1⟩ $\hat{H}$ includes both positive and negative eigenvalues, in this case $\mathcal{R}^*$ is unstable.

⟨2⟩ $\hat{H}$ is negative definite. In this case, the segment of $z^2 < 1$ on $\mathcal{R}^*$ is stable, while the other parts of $\mathcal{R}$ are unstable.

⟨3⟩ $\hat{H}$ is positive definite. In this case, the segments of $z^2 > 1$ on $\mathcal{R}^*$ are stable, while the part of $z^2 < 1$ is unstable.

Proof. First, we calculate the analytical form of Hessian matrix on the opposite singularity.

For the best approximation $\boldsymbol{\xi}^* = (\boldsymbol{v}^*, w^*, \boldsymbol{0}, z \in \mathbb{R})$, $\left.\frac{\partial L(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}}\right|_{\boldsymbol{\xi}=\boldsymbol{\xi}^*} = \boldsymbol{0}$, we can get $\left.\left\langle e(y, \boldsymbol{x}, \boldsymbol{\xi}) \frac{\partial \phi(\boldsymbol{x}, \boldsymbol{v})}{\partial \boldsymbol{v}} \right\rangle\right|_{\boldsymbol{\xi}=\boldsymbol{\xi}^*} = \boldsymbol{0}$.

Let us substitute the best approximation $\boldsymbol{\xi}^*$ into Eq. (38), the Hessian matrix $H(\boldsymbol{\xi})$ on the opposite singularity can be rewritten as:

$$H(\boldsymbol{\xi})|_{\boldsymbol{\xi}=\boldsymbol{\xi}^*} = \begin{bmatrix} F_1(\boldsymbol{\xi}^*) & & \\ & F_2(\boldsymbol{\xi}^*) & \\ & & \boldsymbol{0} \end{bmatrix}, \tag{56}$$

where

$$F_1(\boldsymbol{\xi}^*) = \left.\begin{bmatrix} \frac{\partial^2 L(\boldsymbol{\xi})}{\partial \boldsymbol{v} \partial \boldsymbol{v}^T} & \frac{\partial^2 L(\boldsymbol{\xi})}{\partial \boldsymbol{v} \partial w} \\ \frac{\partial^2 L(\boldsymbol{\xi})}{\partial w \partial \boldsymbol{v}^T} & \frac{\partial^2 L(\boldsymbol{\xi})}{\partial w^2} \end{bmatrix}\right|_{\boldsymbol{\xi}=\boldsymbol{\xi}^*} = \begin{bmatrix} w^{*2} P_4(\boldsymbol{v}^*) - w^* P_5(\boldsymbol{v}^*) & w^* P_2(\boldsymbol{v}^*, \boldsymbol{v}^*) \\ w^* P_2(\boldsymbol{v}^*, \boldsymbol{v}^*)^T & w^* P_1(\boldsymbol{v}^*, \boldsymbol{v}^*) \end{bmatrix}, \tag{57}$$

and

$$F_2(\boldsymbol{\xi}^*) = (z^2 - 1)\hat{H}(\boldsymbol{v}^*, w^*), \tag{58}$$

$$\hat{H}(\boldsymbol{v}^*, w^*) = \frac{1}{4} w^* P_5(\boldsymbol{v}^*), \tag{59}$$

$$P_3(\boldsymbol{t}, \boldsymbol{v}) = -\frac{2}{\pi} \frac{\sqrt{\det(\boldsymbol{B}(\boldsymbol{t}, \boldsymbol{v})^{-1})}}{1+\|\boldsymbol{v}\|^2} \left(\boldsymbol{v}\boldsymbol{t}^T\boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{t}\boldsymbol{v}^T + \boldsymbol{t}^T\boldsymbol{v}\boldsymbol{B}(\boldsymbol{t}, \boldsymbol{v})^{-1}\right), \quad \text{for } \boldsymbol{t} \neq \boldsymbol{v}, \tag{60}$$

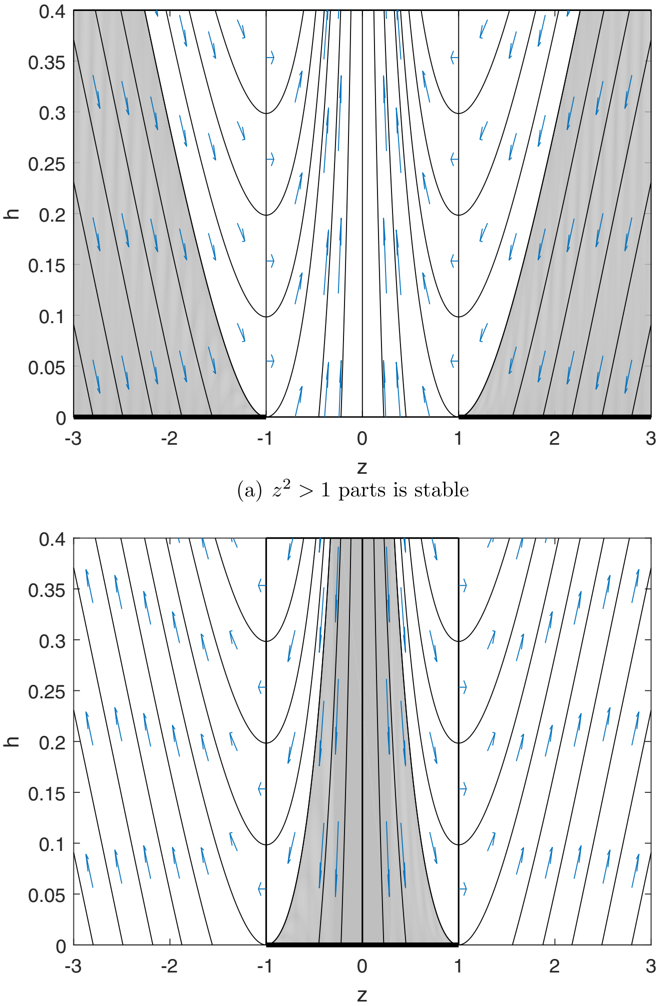
$$P_3(\boldsymbol{v}, \boldsymbol{v}) = \frac{2}{\pi} \frac{\sqrt{\det(\boldsymbol{B}(\boldsymbol{v}, \boldsymbol{v})^{-1})}}{1+\|\boldsymbol{v}\|^2} \left(\boldsymbol{I}_n - \frac{2(2+3\|\boldsymbol{v}\|^2)}{(1+\|\boldsymbol{v}\|^2)(1+2\|\boldsymbol{v}\|^2)} \boldsymbol{v}\boldsymbol{v}^T - (1+\|\boldsymbol{v}\|^2)\boldsymbol{B}(\boldsymbol{v}, \boldsymbol{v})^{-1} \right), \tag{61}$$

$$P_4(\boldsymbol{v}) = \frac{2}{\pi} \sqrt{\det(\boldsymbol{B}(\boldsymbol{v}, \boldsymbol{v})^{-1})} \boldsymbol{B}(\boldsymbol{v}, \boldsymbol{v})^{-1}, \tag{62}$$

$$P_5(\boldsymbol{v}) = v_1 P_3(\boldsymbol{t}_1, \boldsymbol{v}) + v_2 P_3(\boldsymbol{t}_2, \boldsymbol{v}) - w P_3(\boldsymbol{v}, \boldsymbol{v}). \tag{63}$$

The calculation processes of $P_3(\boldsymbol{t}, \boldsymbol{v})$, $P_3(\boldsymbol{v}, \boldsymbol{v})$ and $P_4(\boldsymbol{v})$ are shown in the Appendix. Now the concrete expression of Hessian matrix on the opposite singularity has been obtained.

Then the eigenvalue analysis can be taken. It can be seen that $F_1(\boldsymbol{\xi}^*)$ is the Hessian of $L(\boldsymbol{\xi})$ with respect to $(\boldsymbol{v}, w)$ at $\boldsymbol{\xi} = \boldsymbol{\xi}^*$. Since $(\boldsymbol{v}^*, w^*)$ is the best approximation, namely $(\boldsymbol{v}^*, w^*)$ is the stable point, the eigenvalues of $F_1(\boldsymbol{\xi}^*)$ are all positive. This implies that the stability of the opposite singularity is determined by the eigenvalues of $F_2(\boldsymbol{\xi}^*)$. The stable parts of the opposite singularity are the regions where the eigenvalues of $F_2(\boldsymbol{\xi}^*)$ are all positive. Thus we can obtain Theorem 1. □



(a) $z^2 > 1$ parts is stable

(b) $z^2 < 1$ part is stable

Fig. 2. Theoretical learning trajectories.

From the results in [23], without analyzing the stability of $\mathcal{R}^*$, the theoretical learning trajectories near $\mathcal{R}^*$ are obtained as follows:

$$h = \frac{1}{2}\boldsymbol{u}^T\boldsymbol{u} = \frac{2w^*}{3} \log \frac{(z^2+3)^2}{|z|} + C, \tag{64}$$

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

As we have discussed the stability of $\mathcal{R}^*$, the theoretical learning trajectories near $\mathcal{R}^*$ with stable part and unstable part are shown in Fig. 2. The stable part is determined by the definiteness of $\hat{H}(\boldsymbol{v}^*, w^*)$ by using Theorem 1. For the trajectories $h \sim z$, line $h = 0$ represents the opposite singularity.

4. Simulation experiments

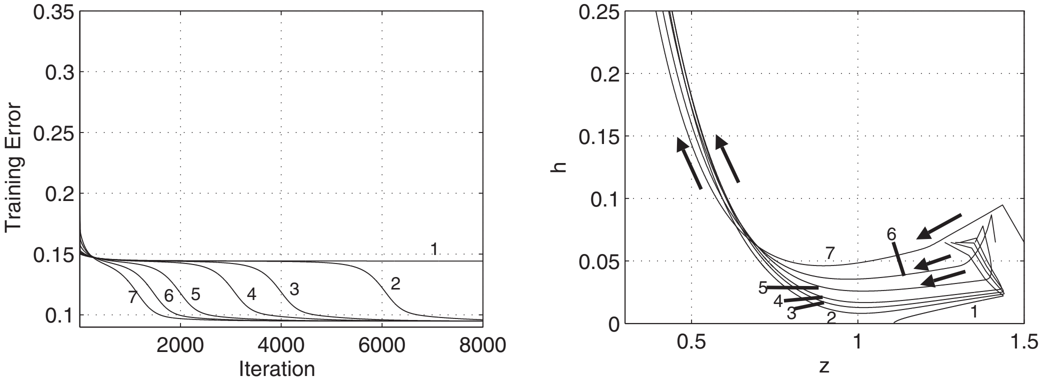
In this section, we verify the theoretical analysis results by taking two simulation experiments.

4.1. Real learning trajectories near the opposite singularity

For a given teacher model, by choosing different initial values of student parameters and using batch mode learning method, we plot the real learning trajectories near the opposite singularity in comparison with theoretical learning trajectories.

Table 1
Initial student parameters of trajectories 1–7 in Fig. 3.

Trajectory	Initial weight $\boldsymbol{J}_1$	Initial weight w_1	Initial weight $\boldsymbol{J}_2$	Initial weight w_2
1	$(0.40,\ 0.10)^T$	−0.12	$(-0.49, 0.35)^T$	−1.04
2	$(0.40,\ 0.10)^T$	−0.12	$(-0.49, 0.35)^T$	−0.98
3	$(0.40,\ 0.10)^T$	−0.12	$(-0.49, 0.35)^T$	−0.94
4	$(0.40,\ 0.10)^T$	−0.12	$(-0.49, 0.35)^T$	−0.90
5	$(0.40,\ 0.10)^T$	−0.12	$(-0.49, 0.35)^T$	−0.80
6	$(0.40,\ 0.10)^T$	−0.12	$(-0.49, 0.35)^T$	−0.70
7	$(0.40,\ 0.10)^T$	−0.12	$(-0.49, 0.35)^T$	−0.60



(a) Trajectories of training error

(b) $h \sim z$ trajectories

Fig. 3. Real batch mode learning dynamics.

The chosen teacher model is:

$$f_0(\boldsymbol{x}, \boldsymbol{\theta}_0) = v_1\phi(\boldsymbol{x}, \boldsymbol{t}_1) + v_2\phi(\boldsymbol{x}, \boldsymbol{t}_2), \tag{65}$$

where $\boldsymbol{t}_1 = [0.2,\ -0.5]^T,\ v_1 = 0.5,\ \boldsymbol{t}_2 = [0.7,\ 0.4]^T,\ v_2 = 0.3$.

We use batch mode learning method to accomplish the experiment. We generate 200 training examples which are subject to the standard Gaussian distribution and the additive noise is subject to Gaussian distribution with mean 0 and variance 0.05. For a chosen learning rate $\eta = 0.005$, the model is trained by the standard gradient algorithm for 8000 times.

By letting the initial states of the two student units be opposite, we get the best approximation $\boldsymbol{v}^* = [0.3367,\ -0.1697]^T$ and $w^* = 0.7984$. When we substitute $\boldsymbol{v}^*$, w^* into Eqs. (33) and (34), we can see that ($\boldsymbol{v}^*$, w^*) is indeed the solution of the two equations. This verifies the correctness of the expression of best approximation.

Then we calculate $\hat{H}(\boldsymbol{v}^*, w^*)$:

$$\hat{H}(\boldsymbol{v}^*, w^*) = \begin{bmatrix} 0.0003 & 0.0007 \\ 0.0007 & 0.0015 \end{bmatrix}. \tag{66}$$

The eigenvalues of $\hat{H}(\boldsymbol{v}^*, w^*)$ is 0.00002 and 0.0018, respectively, which implies $\hat{H}(\boldsymbol{v}^*, w^*)$ is a positive definite matrix. According to Theorem 1, the segments of $z^2 > 1$ on $\mathcal{R}$ are stable, while the part of $z^2 < 1$ is unstable, which means the learning dynamics near opposite singularity would be in accordance with Fig. 2(a).

Then we try different initial student parameters and plot the trajectories of training error and $h \sim z$ in comparison with the trajectories in Fig. 2(a). The initial student parameters are shown in Table 1 and we only make a slight change in the initial parameters. The results are shown in Fig. 3. The trajectories $'1' \sim '7'$ correspond to the initial states of student model take values $'1' \sim '7'$ in Table 1, respectively.

As shown in Fig. 3(b), it can be seen that the $z > 1$ part of $h \sim z$ trajectories near $\mathcal{R}^*$ are attractive, namely the $z^2 > 1$ part is stable and $z^2 < 1$ part is unstable. The trajectories are very similar to those in the right part of Fig. 2(a). This is in accordance with the above analysis of definite of $H(\boldsymbol{v}^*,\ w^*)$. As the theoretical learning trajectories are obtained by using Taylor expansion at $h = 0$, the difference between the real batch mode learning trajectory and the analytical learning trajectory becomes significant with the increase of h.

4.2. Cuff-less blood pressure estimation

In the prior experiment, we have shown the real learning trajectories in accordance with the theoretical analysis. In this section, we do a real experiment to verify the obtained results. In the Physionets multiparameter Intelligent Monitoring in Intensive Care (MIMIC) II (version 3, accessed on September 2015) online waveform database [24], the arterial blood pressure (ABP) signal is estimated by the photoplethysmograph (PPG) and electrocardiogram (ECG) signal, namely for the learning machine, the input is $\boldsymbol{x} = [x_1,\ x_2]^T$ and the output is y, where x_1 is PPG, x_2 is ECG and y is ABP. In order to obtain better simulation results, the Gaussian normalization is chosen to perform the preprocessing. $\boldsymbol{x}(k)$ is normalized as:

$$\boldsymbol{x}(k) = \frac{\boldsymbol{x}(k) - \boldsymbol{\mu}}{\boldsymbol{\delta}}, \tag{67}$$

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

$$\boldsymbol{\mu} = \frac{1}{M}\sum_{i=1}^{M}\boldsymbol{x}(i), \tag{68}$$

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

$$\boldsymbol{\delta} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(\boldsymbol{x}(i) - \boldsymbol{\mu})^2}, \tag{69}$$

and M is the sample number of the data set.

We use the batch mode learning to accomplish the experiment. The number of hidden units in the student model is chosen to be

Table 2
Initial student parameters belong to $z^2 < 1$ part.

Initial state of $\boldsymbol{J}$	$\boldsymbol{J}^{(0)} = \begin{bmatrix} -0.9917 & -0.9818 & -0.8861 & 0.5916 & -0.0808 & -0.6379 \\ 0.1070 & -0.2837 & -0.3376 & -0.9780 & -0.6434 & -0.8802 \end{bmatrix}$
Initial state of $\boldsymbol{w}$	$\boldsymbol{w}^{(0)} = [-0.6461, -0.7087, -0.2742, 0.6097, -0.8106, -0.9708]$
Final state of $\boldsymbol{J}$	$\boldsymbol{J} = \begin{bmatrix} -1.3827 & -1.7233 & -2.7797 & 1.3876 & 2.2159 & -0.4898 \\ 1.0562 & -1.4124 & -2.1164 & -1.0765 & -1.2228 & -1.1249 \end{bmatrix}$
Final state of $\boldsymbol{w}$	$\boldsymbol{w} = [-0.7850, -1.0321, 0.8594, 0.8656, -1.7105, 0.1794]$

Table 3
Initial student parameters belong to $z^2 > 1$ part.

Initial state of $\boldsymbol{J}$	$\boldsymbol{J}^{(0)} = \begin{bmatrix} -0.9917 & -0.9818 & -0.8861 & 0.5916 & -0.0808 & -0.6379 \\ 0.1070 & -0.2837 & -0.3376 & -0.9780 & -0.6434 & -0.8802 \end{bmatrix}$
Initial state of $\boldsymbol{w}$	$\boldsymbol{w}^{(0)} = [0.5761, -0.7087, -0.2742, 0.3097, -0.8106, -0.9708]$
Final state of $\boldsymbol{J}$	$\boldsymbol{J} = \begin{bmatrix} -2.2708 & -1.2955 & -2.3341 & -0.1449 & 0.3998 & -1.2210 \\ 0.9941 & 0.7223 & -1.5141 & -0.9618 & -0.6218 & -0.8934 \end{bmatrix}$
Final state of $\boldsymbol{w}$	$\boldsymbol{w} = [1.6272, -1.9726, 0.7033, 0.4782, -0.7119, -0.8947]$

Table 4
Simulation results of blood pressure estimation.

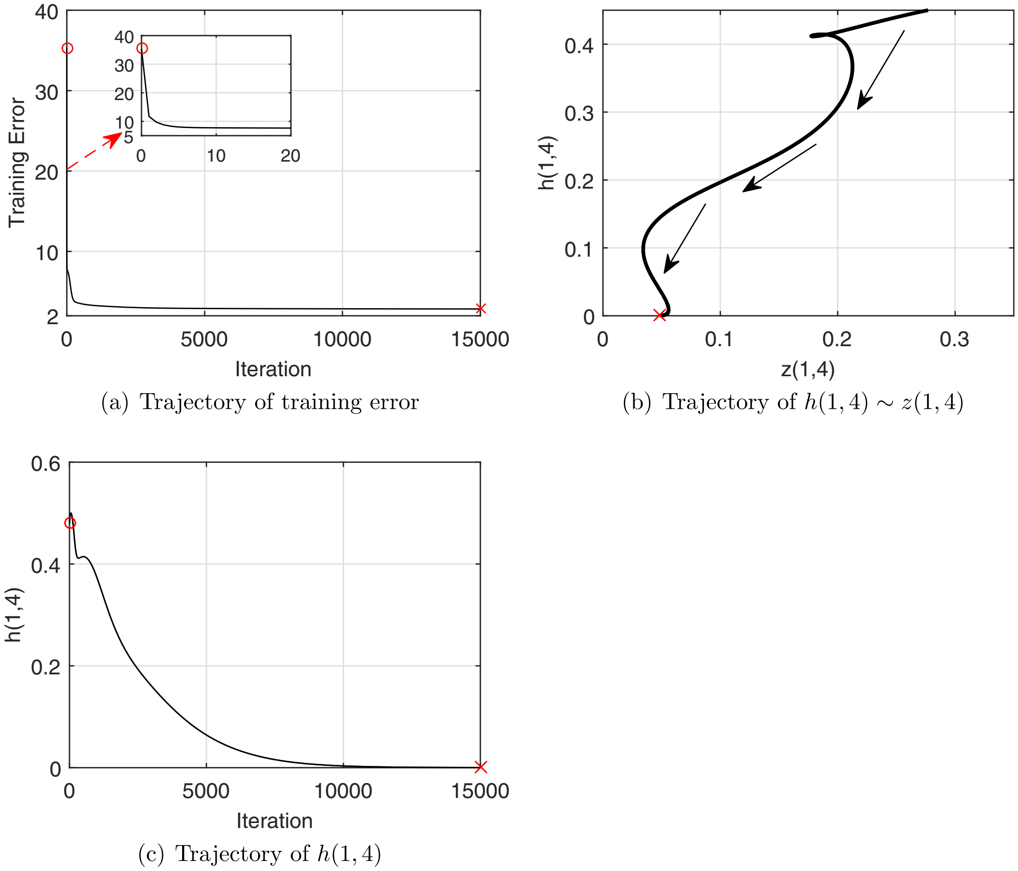
	Training error	Test error
Initial student parameters belong to $z^2 < 1$ part	2.8407	7.3947
Initial student parameters belong to $z^2 > 1$ part	2.7005	6.4823

$k = 6$, namely the student MLP is given by:

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{i=1}^{6} w_i \phi(\boldsymbol{x}, \boldsymbol{J}_i). \qquad (70)$$

We use $N = 200$ samples to train the MLPs. For the learning rate $\eta = 0.03$, the model is trained by the typical backpropagation (BP) algorithm for 15000 times. After training, we use another 200 samples to test the training efficiency and choose the root mean square error (RMSE), $\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$, to measure the test error. $h(i, j) = \frac{1}{2}(\boldsymbol{J}_i + \boldsymbol{J}_j)^T(\boldsymbol{J}_i + \boldsymbol{J}_j)$ and $z(i, j) = \frac{w_j + w_i}{w_j - w_i}$ have the same meanings with h and z in Eqs. (64) and (16), respectively.

In this section, we give two examples that the student parameters arrive in the stable part and unstable part of the opposite singularity, respectively. The initial student parameters of the two example are exactly the same except for $w_1^{(0)}$ and $w_4^{(0)}$, where the

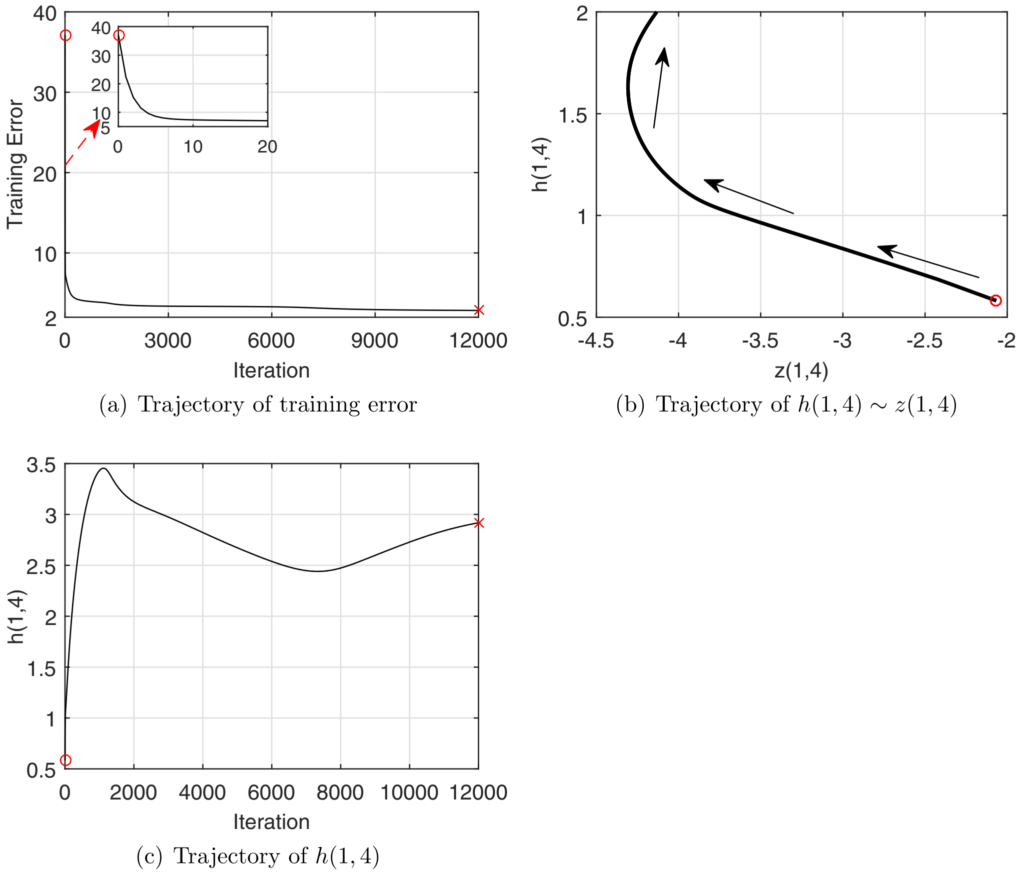


(a) Trajectory of training error

(b) Trajectory of $h(1, 4) \sim z(1, 4)$

(c) Trajectory of $h(1, 4)$

Fig. 4. Initial student parameters belong to $z^2 < 1$ part.



(a) Trajectory of training error

(b) Trajectory of $h(1,4) \sim z(1,4)$

(c) Trajectory of $h(1,4)$

Fig. 5. Initial student parameters belong to $z^2 > 1$ part.

initial values of the two examples are in the $z^2 < 1$ part and $z^2 > 1$ part, respectively. The initial values and final values of the student parameters of the two examples are shown in Tables 2 and 3, respectively. The training error and test error are shown in Table 4.

For the case that the student parameters arrive in the stable part of the opposite singularity, as shown in Fig. 2, the learning process will be trapped in the opposite singularity and finally can not escape from it. The simulation results are given in Fig. 4, which show the trajectories of the training error, $h(1,\ 4) \sim z(1,\ 4)$ and $h(1,\ 4)$, respectively. In Fig. 4, '$\circ$' and '$\times$' represent the initial state and final state, respectively.

For the case that the student parameters arrive in the unstable part of the opposite singularity, the learning process will be affected by the opposite singularity slightly. The simulation results are given in Fig. 5, which show the trajectories of the training error, $h(1,\ 4) \sim z(1,\ 4)$ and $h(1,\ 4)$, respectively. In Fig. 5, '$\circ$' and '$\times$' represent the initial state and final state, respectively.

From Fig. 4(c), after the training process $h(1,\ 4)$ nearly equals 0, the final state $\boldsymbol{J}_1 = [-1.3827,\ \ 1.0562]^T$ and $\boldsymbol{J}_4 = [1.3876,\ \ -1.0765]^T$, $\boldsymbol{J}_1$ nearly equals to $-\boldsymbol{J}_4$, which means that the hidden nodes 1 and 4 have arrived in the opposite singularity. Meanwhile, $h(1,\ 4) \sim z(1,\ 4)$ curve shown in Fig. 4(b) in similar to the gray part (namely the stable part) in Fig. 2(b), thus $z^2 < 1$ part of the opposite singularity is stable. As shown in Fig. 5(b), the $h(1,\ 4) \sim z(1,\ 4)$ curve is very similar to that in the left part of Fig. 2(b), which implies that the $z^2 > 1$ part of the opposite singularity is unstable.

Though we can not obtain the definiteness of $\hat{H}(\boldsymbol{v}^*, w^*)$ due to the unknown teacher parameters, the results in Figs. 4(b) and 5(b) indicate that $z^2 < 1$ part of the opposite singularity is stable and the other part is unstable. This verity the stability analysis results obtained in Theorem 1. From the results in Table 4, we can see that the training error and test error belong to $z^2 < 1$ part are bigger than those belong to $z^2 > 1$ part. This is because $z^2 < 1$ part of the opposite singularity is stable and the learning process belongs to this case will be trapped in the opposite singularity and converge to a local minimum, not the global minimum. Thus it is important to avoid the stable part of the opposite singularity to reduce the influence of the singularity.

5. Conclusion

The multilayer perceptrons (MLPs) are widely used in many fields. However, the learning processes of MLPs have some strange behaviors, such as the learning behavior maybe become very slow and MLPs often exhibit a plateau phase. This is mainly caused by the existed singularities in the parameter space of MLPs. For the bipolar-activation-function multilayer perceptrons, the parameter space has opposite singular regions, where the Fisher information matrix degenerates. As stability plays a significant role in analyzing the properties of the opposite singularity, in this paper, we take the stability analysis of the opposite singularity of MLPs by analyzing the eigenvalues of the Hessian matrix based on loss function. After giving the explicit expressions of $P_3(\boldsymbol{t},\ \boldsymbol{v})$ and $P_4(\boldsymbol{v})$, we first give the analytical form of best approximation, and then we obtain the concrete expression of Hessian matrix on the opposite singularity. Further, the stability of the opposite singularity is investigated and the learning dynamics near opposite singularity are discussed. In the simulation part, the results obtained in this paper are verified by taking two experiments.

Acknowledgments

This project is supported by Major Program of National Natural Science Foundation of China under grant 11190015 and National Natural Science Foundation of China under grant 61374006 and Shandong Province Natural Science Foundation under grant ZR2016FL12 and Project of Shandong Province Higher Educational Science and Technology Program under grant J16L117 and Jiangsu Planned Projects for Postdoctoral Research Funds.

Appendix

For $\boldsymbol{t} \neq \boldsymbol{v}$,

$$\begin{aligned}
P_3(\boldsymbol{t}, \boldsymbol{v}) &= \left\langle \phi(\boldsymbol{x}, \boldsymbol{t}) \frac{\partial^2 \phi(\boldsymbol{x}, \boldsymbol{v})}{\partial \boldsymbol{v} \partial \boldsymbol{v}^T} \right\rangle \\
&= \frac{\partial}{\partial \boldsymbol{v}^T} \left\langle \phi(\boldsymbol{x}, \boldsymbol{t}) \frac{\partial \phi(\boldsymbol{x}, \boldsymbol{v})}{\partial \boldsymbol{v}} \right\rangle = \frac{\partial}{\partial \boldsymbol{v}^T} P_2(\boldsymbol{t}, \boldsymbol{v}) \\
&= -\frac{2}{\pi} \frac{\sqrt{\det(\boldsymbol{B}(\boldsymbol{t}, \boldsymbol{v})^{-1})}}{1 + \|\boldsymbol{v}\|^2} \left(\boldsymbol{v}\boldsymbol{t}^T \boldsymbol{A}^{-1} + \boldsymbol{A}^{-1} \boldsymbol{t} \boldsymbol{v}^T + \boldsymbol{t}^T \boldsymbol{v} \boldsymbol{B}(\boldsymbol{t}, \boldsymbol{v})^{-1} \right).
\end{aligned} \tag{A-1}$$

For $P_2(\boldsymbol{v}, \boldsymbol{v})$, we have:

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{v}^T} P_2(\boldsymbol{v}, \boldsymbol{v}) &= \left\langle \frac{\partial \phi(\boldsymbol{x}, \boldsymbol{v})}{\partial \boldsymbol{v}^T} \frac{\partial^2 \phi(\boldsymbol{x}, \boldsymbol{v})}{\partial \boldsymbol{v}} \right\rangle + \left\langle \phi(\boldsymbol{x}, \boldsymbol{v}) \frac{\partial^2 \phi(\boldsymbol{x}, \boldsymbol{v})}{\partial \boldsymbol{v} \partial \boldsymbol{v}^T} \right\rangle \\
&= P_4(\boldsymbol{v}) + P_3(\boldsymbol{v}, \boldsymbol{v}),
\end{aligned} \tag{A-2}$$

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{v}^T} P_2(\boldsymbol{v}, \boldsymbol{v}) &= \frac{\partial}{\partial \boldsymbol{v}^T} \left(\frac{2}{\pi} \sqrt{\det(\boldsymbol{B}(\boldsymbol{v}, \boldsymbol{v})^{-1})} \boldsymbol{A}^{-1} \boldsymbol{t} \right) \\
&= \frac{2}{\pi} \frac{\sqrt{\det(\boldsymbol{B}(\boldsymbol{v}, \boldsymbol{v})^{-1})}}{1 + \|\boldsymbol{v}\|^2} \\
&\quad \left(\boldsymbol{I}_n - \frac{2(2 + 3\|\boldsymbol{v}\|^2)}{(1 + \|\boldsymbol{v}\|^2)(1 + 2\|\boldsymbol{v}\|^2)} \boldsymbol{v}\boldsymbol{v}^T \right).
\end{aligned} \tag{A-3}$$

Now we calculate $P_4(\boldsymbol{v})$. From Eq. (1), we have:

$$y - f_0(\boldsymbol{x}) = \varepsilon \sim \mathcal{N}(0, 1), \tag{A-4}$$

then

$$\begin{aligned}
&\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp(-\frac{1}{2}(y - f_0(\boldsymbol{x}))^2) \mathrm{d}y \\
&\quad = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp(-\frac{\varepsilon^2}{2}) \mathrm{d}\varepsilon = 1.
\end{aligned} \tag{A-5}$$

$P_4(\boldsymbol{v})$ can be rewritten as:

$$\begin{aligned}
P_4(\boldsymbol{v}) =& (2\pi)^{-\frac{n}{2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{\partial \phi(\boldsymbol{x}, \boldsymbol{v})}{\partial \boldsymbol{v}} \frac{\partial \phi(\boldsymbol{x}, \boldsymbol{v})}{\partial \boldsymbol{v}^T} \exp(-\frac{1}{2}\|\boldsymbol{x}\|^2) \\
&\times \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y - f_0(\boldsymbol{x}))^2 \right) \mathrm{d}y \mathrm{d}\boldsymbol{x} \\
=& (2\pi)^{-\frac{n}{2}} \int_{-\infty}^{\infty} \frac{\partial \phi(\boldsymbol{x}, \boldsymbol{v})}{\partial \boldsymbol{v}} \frac{\partial \phi(\boldsymbol{x}, \boldsymbol{v})}{\partial \boldsymbol{v}^T} \exp(-\frac{1}{2}\|\boldsymbol{x}\|^2) \mathrm{d}\boldsymbol{x} \\
=& \frac{2}{\pi} (\sqrt{2\pi})^{-n} \int_{-\infty}^{\infty} \boldsymbol{x}\boldsymbol{x}^T \exp\left(-\frac{1}{2} \left(\boldsymbol{x}^T (\boldsymbol{I}_n + 2\boldsymbol{v}\boldsymbol{v}^T) \boldsymbol{x} \right) \right) \mathrm{d}\boldsymbol{x} \\
=& \frac{2}{\pi} \sqrt{\det(\boldsymbol{B}(\boldsymbol{v}, \boldsymbol{v})^{-1})} \boldsymbol{B}(\boldsymbol{v}, \boldsymbol{v})^{-1}.
\end{aligned} \tag{A-6}$$

Then $P_3(\boldsymbol{v}, \boldsymbol{v})$ can also be obtained according to formula (A-2):

$$\begin{aligned}
P_3(\boldsymbol{v}, \boldsymbol{v}) &= \frac{\partial}{\partial \boldsymbol{v}^T} P_2(\boldsymbol{v}, \boldsymbol{v}) - P_4(\boldsymbol{v}) \\
&= \frac{2}{\pi} \frac{\sqrt{\det(\boldsymbol{B}(\boldsymbol{v}, \boldsymbol{v})^{-1})}}{1 + \|\boldsymbol{v}\|^2} \left(\boldsymbol{I}_n - \frac{2(2 + 3\|\boldsymbol{v}\|^2)}{(1 + \|\boldsymbol{v}\|^2)(1 + 2\|\boldsymbol{v}\|^2)} \boldsymbol{v}\boldsymbol{v}^T \right. \\
&\quad \left. - (1 + \|\boldsymbol{v}\|^2) \boldsymbol{B}(\boldsymbol{v}, \boldsymbol{v})^{-1} \right).
\end{aligned} \tag{A-7}$$

References

[1] C. Mu, Z. Ni, C. Sun, H. He, Data-driven tracking control with adaptive dynamic programming for a class of continuous-time nonlinear systems, IEEE Trans. Cybern. 47 (6) (2017) 1460–1470.

[2] C. Mu, Z. Ni, C. Sun, H. He, Air-breathing hypersonic vehicle tracking control based on adaptive dynamic programming, IEEE Trans. Neural Netw. Learn. Syst. 28 (3) (2017) 584–598.

[3] Z. Zheng, L. Sun, Path following control for marine surface vessel with uncertainties and input saturation, Neurocomputing 177 (2016) 158–167.

[4] S. Amari, H. Nagaoka, Information Geometry, AMS and Oxford University Press, New York, 2000.

[5] S. Amari, T. Ozeki, Differential and algebraic geometry of multilayer perceptrons, IEICE Trans. Fundam. Electron. Commun. Comput. Sci. E84-A (2001) 31–38.

[6] D. Saad, A. Solla, Exact solution for online learning in multilayer neural networks, Phys. Rev. Lett. 74 (21) (1995) 4337–4340.

[7] M. Biehl, H. Schwarze, Learning by on-line gradient descent, J. Phys. A Math. Gen. 28 (3) (1995) 643–656.

[8] S. Amari, T. Ozeki, F. Cousseau, H. Wei, Dynamics of learning in hierarchical models - singularity and milnor attractor, The proceedings of the Second International Conference on Cognitive Neurodynamics, 2009, pp. 3–9.

[9] F. Cousseau, T. Ozeki, S. Amari, Dynamics of learning in multilayer perceptrons near singularities, IEEE Trans. Neural Netw. 19 (8) (2008) 1313–1328.

[10] H. Park, M. Inoue, M. Okada, Online learning dynamics of multilayer perceptrons with unidentifiable parameters, J. Phys. A Math. Gen. 36 (47) (2003) 11753–11764.

[11] S. Amari, H. Park, T. Ozeki, Singularities affect dynamics of learning in neuromanifolds, Neural Comput. 18 (5) (2006) 1007–1065.

[12] S. Amari, Natural gradient works efficiently in learning, Neural Comput. 10 (2) (1998) 251–276.

[13] J. Zhao, H. Wei, C. Zhang, W. Li, W. Guo, K. Zhang, Natural gradient learning algorithms for RBF networks, Neural Comput. 27 (2) (2015) 481–505.

[14] S. Watanabe, Algebraic analysis for non-identifiable learning machines, Neural Comput. 13 (4) (2001) 899–933.

[15] S. Watanabe, Algebraic geometrical methods for hierarchical learning machines, Neural Netw. 14 (8) (2001) 1049–1060.

[16] S. Watanabe, A widely applicable Bayesian information criterion., J. Mach. Learn. Res. 14 (2013) 867–897.

[17] K. Fukumizu, S. Amari, Local minima and plateaus in hierarchical structure of multilayer perceptrons, Neural Netw. 13 (3) (2000) 317–327.

[18] W. Guo, H. Wei, J. Zhao, K. Zhang, Averaged learning equations of error–function-based multilayer perceptrons, Neural Comput. Appl. 25 (3–4) (2014) 825–832.

[19] W. Guo, H. Wei, J. Zhao, K. Zhang, Theoretical and numerical analysis of learning dynamics near singularity in multilayer perceptrons, Neurocomputing 151 (2015) 390–400.

[20] H. Wei, J. Zhang, F. Cousseau, T. Ozeki, S. Amari, Dynamics of learning near singularities in layered networks, Neural Comput. 20 (3) (2008) 813–843.

[21] H. Wei, S. Amari, Dynamics of learning near singularities in radial basis function networks, Neural Netw. 21 (7) (2008) 989–1005.

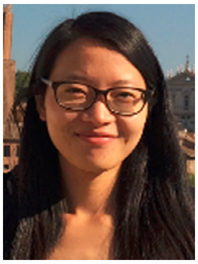
[22] H. Park, T. Ozeki, Singularity and slow convergence of the EM algorithm for gaussian mixtures, Neural Process. Lett. 29 (1) (2009) 45–59.

[23] W. Guo, H. Wei, J. Zhao, K. Zhang, Theoretical analysis of learning dynamics near the opposite singularities in multilayer perceptrons, Control Theory Appl. 31 (2) (2014) 140–147. (in Chinese)

[24] M. Kachuee, M.M. Kiani, H. Mohammadzade, M. Shabany, Cuff-less high-accuracy calibration-free blood pressure estimation using pulse transit time, in: IEEE International Symposium on Circuits and Systems (ISCAS'15), 2015, pp. 1006–1009.

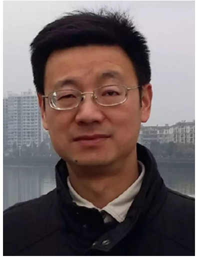
Weili Guo was born in Jining, China, in 1987. He received the B.S. degree in School of Science, Shandong Jianzhu University, China in 2007, the M.S. degree in School of Science, Nanjing Agricultural University, China in 2010, and the Ph.D. degree in School of Automation, Southeast University, China in 2014. He is currently a postdoctoral in Southeast University, China. His main research is in singular learning dynamics of feedforward neural networks and deep neural networks.

Junsheng Zhao received the B.S. degree in the Department of Mathematics, Liaocheng University, China in 2003, the M.S. degree in Qufu Normal University, China in 2006 and the Ph.D. degree in Southeast University, China in 2015. He is currently a lecturer in the School of Mathematical Sciences, Liaocheng University, China. His research interests include real and artificial in neural networks and identification of nonlinear systems.

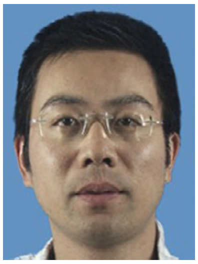


Jinxia Zhang received the B.S. degree in the Department of Computer Science and Engineering, Nanjing University of Science and Technology, China in 2009 and the Ph.D. degree in the Department of Computer Science and Engineering, Nanjing University of Science and Technology, China in 2015. She was a visiting scholar in Visual Attention Lab, Brigham and Women's Hospital from 2012 to 2014. She is currently a lecturer in the School of Automation, Southeast University. Her research interests include visual attention, visual saliency detection, computer vision and machine learning

Haikun Wei received the B.S. degree in the Department of Automation, North China University of Technology, China in 1994, and the M.S. and Ph.D. degrees in the Research Institute of Automation, Southeast University, China in 1997 and 2000. He was a visiting scholar in RIKEN Brain Science Institute, Japan from 2005 to 2007. His research interests include real and artificial in neural networks and industry automation.



Aiguo Song received the B.S. degree in automatic control in 1990, the M.S. degree in measurement and control in 1993 from Nanjing Aeronautics and Astronautics University, Nanjing, China, and the Ph.D. degree in measurement and control from Southeast University, Nanjing, China, 1996. He is currently a professor and director of Robot Sensor and Control Laboratory, and Dean of School of Instrument Science and Engineering, Southeast University, China. 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.



Kanjian Zhang received the B.S. degree in mathematics from Nankai University, China in 1994, and the M.S. and Ph.D. degrees in control theory and control engineering from Southeast University, China in 1997 and 2000. He is currently a professor in the School of Automation, Southeast University. His research is in nonlinear control theory and its applications, with particular interest in robust output feedback design and optimization control.