Statistical mechanical analysis of learning dynamics of two-layer perceptron with multiple output units

OPEN ACCESS
IOP Publishing
Journal of Physics A: Mathematical and Theoretical
J. Phys. A: Math. Theor. **52** (2019) 184002 (17pp)
https://doi.org/10.1088/1751-8121/ab0669

Statistical mechanical analysis of learning dynamics of two-layer perceptron with multiple output units

Yuki Yoshida[1], Ryo Karakida[2], Masato Okada[1,2,3,4] and Shun-Ichi Amari[3]

[1] Department of Complexity Science and Engineering, Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8561, Japan
[2] Artificial Intelligence Research Center, AIST, Koto, Tokyo 135-0064, Japan
[3] RIKEN Brain Science Institute, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan

E-mail: okada@edu.k.u-tokyo.ac.jp

Received 16 July 2018, revised 25 December 2018
Accepted for publication 12 February 2019
Published 3 April 2019

Abstract

The plateau phenomenon, wherein the loss value stops decreasing during the process of learning, is troubling. Various studies suggest that the plateau phenomenon is frequently caused by the network being trapped in the singular region on the loss surface, a region that stems from the symmetrical structure of neural networks. However, these studies all deal with networks that have a one-dimensional output, and networks with a multidimensional output are overlooked. This paper uses a statistical mechanical formalization to analyze the dynamics of learning in a two-layer perceptron with multidimensional output. We derive order parameters that capture macroscopic characteristics of connection weights and the differential equations that they follow. We show that singular-region-driven plateaus diminish or vanish with multidimensional output, in a simple setting. We found that the more non-degenerative (i.e. far from one-dimensional output) the model is, the more plateaus are alleviated. Furthermore, we showed theoretically that singular-region-driven plateaus seldom occur in the learning process in the case of orthogonalized initializations.

Keywords: neural network, plateau, singular region, statistical mechanics

S Supplementary material for this article is available online

(Some figures may appear in colour only in the online journal)

[4] Author to whom any correspondence should be addressed.

1. Introduction

Since deep learning was proposed in the late 2000s, neural networks have received much attention. That is because they enabled us to solve real-world tasks in various fields, including image recognition, speech recognition and machine translation tasks, with performance far exceeding conventional methods. However, there are some problems with the neural networks left behind. One of them is the 'plateau phenomenon', the main topic of this study, which we describe in detail below.

The perceptron is one representative of machine learning methods. Although it was first proposed in 1958 [1], there was no efficient learning algorithm at that time, and it became obsolete. In 1985, with discovery of the backpropagation method [2], which is a fundamental learning algorithm of neural networks, neural networks began to gain attention once again. However, another problem occurred; that is the 'plateau phenomenon', wherein the learning slows down partway through. In the learning process of neural networks, weight parameters of the neural network are updated iteratively so that the loss (gap between the network's output and desired output) decreases. However, typically the loss does not decrease simply, but its decreasing speed slows down significantly partway through learning, and then it speeds up again after a long period of time (see the black line in figure 2(a) and figure S1 in the supplemental material (stacks.iop.org/JPhysA/52/184002/mmedia) for example). This is what we call the 'plateau phenomenon'. The phenomenon is observed ubiquitously in the learning of hierarchical models, including neural networks, radial basis function networks and a mixture of expert models [3–10]. However, in recent years, although many researchers and engineers train hierarchical neural networks, the plateau phenomenon is rarely observed in practical applications of deep learning. Why is that?

With regard to theoretical studies of learning dynamics, that of linear neural networks have been studied analytically [11–13]. However, the plateau phenomenon is specific to nonlinear neural network, which has a nonlinear activation function. Although studying the learning dynamics of nonlinear neural networks is challenging, the underlying mechanism of the plateau phenomenon has been widely studied. It is known empirically that neural networks get trapped into a plateau because they have symmetrical weights such that their input–output relationship is invariant under swapping two units of a hidden layer. The learning dynamics of neural networks have been studied in various settings. Some past studies, using statistical mechanical formalization, derived the learning dynamics of a two-layer soft committee machine [4] and a two-layer perceptron [3]; this research showed that the weight symmetry results in saddle points that cause plateauing. Furthermore, it has been recognized that such a plateau phenomenon stems from a singular structure in the parameter space (see [9, 14–16], or [17] for a review). In the parameter space, a Riemannian metric is naturally induced by a Fisher information matrix, which represents how two models identified by slightly different parameters differ as statistical models [18]. This metric is not necessarily regular, but there are regions in the parameter space in which the metric degenerates, called singular regions [14]. More specifically, when we consider a two-layer neural network, if it has two identical weight vectors projected from the input layer to two different hidden units, its input–output relationship can be realized by another model which has one lesser hidden units than the original model. If this downsized model gives local minima of the loss in the downsized parameter space, the original model parameter is in the singular region (see also the section 'Singular regions and plateaus' for detailed explanation). The gradient of the loss is zero everywhere within the singular region. Although an isolated saddle point has the same property, the singular region and a saddle point is different to an isolated saddle point in two ways; the singular region has one-or-more dimensions, and it is a Milnor-like attractor [19], that is, it has a

region of attraction which has nonzero measure (see figure 5 in [14] for a schematic diagram of the singular region). However, all of this research assumes one-dimensional output; in other words, networks that have multiple output units are overlooked, and they may avoid the plateau phenomena, which is a usual situation in modern deep learning and seems to be rational.

For these reasons, we analyze the learning dynamics of a two-layer nonlinear perceptron that has multiple output units, with statistical mechanical formulation. First, we introduce the student–teacher learning setting for ease of analysis [4, 20]. Second, we introduce several order parameters, which can capture macroscopic characteristics of network weights, and derive their evolution equations from that of microscopic variables (i.e. network weights). We solve derived equations numerically to obtain macroscopic learning dynamics, with which we can discuss plateau phenomenon quantitatively. Under a simple setting, we show that singular-region-driven plateaus diminish or vanish with multidimensional output. We find that the less degenerative (i.e. like one-dimensional output) the model is, the less the model approaches the singular region, and then the more the plateau shortens. Further, we show theoretically that singular-region-driven plateaus seldom occur in the learning process if the student and teacher models are initialized orthogonally.

Note that the claim that having multiple output units might alleviate plateau phenomenon is hypothesized by intuitive insight in our previous work [21], but in the current work we examine the hypothesis by experiments and theoretical analyses and show that multiple output units can indeed prevent approaching the singular region and vanish plateaus.

2. Model

We considered a neural network with an input layer (size N), a hidden layer (size K), and an output layer (size O). The network receives input data $\boldsymbol{\xi} \in \mathbb{R}^N$ and calculates output $\boldsymbol{s} = \sum_{i=1}^{K} \boldsymbol{w}_i g(\boldsymbol{J}_i \cdot \boldsymbol{\xi}) \in \mathbb{R}^O$, where g is the activation function. The parameters $\boldsymbol{J}$ and $\boldsymbol{w}$ are optimized during learning depending on the difference between the output $\boldsymbol{s}$ and the teacher data $\boldsymbol{t}$. We considered an ideal situation in which the teacher data $\boldsymbol{t}$ is determined as $\boldsymbol{t} = \sum_{n=1}^{M} \boldsymbol{v}_n g(\boldsymbol{B}_n \cdot \boldsymbol{\xi}) \in \mathbb{R}^O$; in other words, the learning network (the 'student network') learns the input–output relationship of the 'teacher network', which is also a two-layer network and has N-M-O units and original fixed weights $\boldsymbol{B}$ and $\boldsymbol{v}$ (figure 1(a)). We assumed the squared loss function $\varepsilon = \frac{1}{2}\|\boldsymbol{s} - \boldsymbol{t}\|^2$, which is most commonly used for regression.

For the statistical mechanical formulation of online learning, we introduced further idealization. We assumed that the dimension of input data N is very large and that each element of input data $\boldsymbol{\xi}$ is generated in accordance with i.i.d. normal distribution, $\mathcal{N}(\xi_i|0,1)$. We put $x_i := \boldsymbol{J}_i \cdot \boldsymbol{\xi}$ and $y_n := \boldsymbol{B}_n \cdot \boldsymbol{\xi}$ and define $Q_{ij} := \boldsymbol{J}_i \cdot \boldsymbol{J}_j = \langle x_i x_j \rangle$, $R_{in} := \boldsymbol{J}_i \cdot \boldsymbol{B}_n = \langle x_i y_n \rangle$, $T_{nm} := \boldsymbol{B}_n \cdot \boldsymbol{B}_m = \langle y_n y_m \rangle$ and $D_{ij} := \boldsymbol{w}_i \cdot \boldsymbol{w}_j$, $E_{in} := \boldsymbol{w}_i \cdot \boldsymbol{v}_n$, $F_{nm} := \boldsymbol{v}_n \cdot \boldsymbol{v}_m$.

The parameters Q_{ij}, R_{in}, T_{nm}, D_{ij}, E_{in}, and F_{nm} defined above capture the state of the system macroscopically; they are therefore called 'order parameters'. The first three represent the state of the first layers of the two networks (student and teacher), and the latter three represent their second layers' state (figure 1(b)). Roughly speaking, Q represents the norm of the student's first layer and T represents that of the teacher's first layer. R is related to similarity between the student and teacher's first layer. D, E, F is the second layers' counterpart of Q, R, T. Among these six order parameter matrices, the values of Q_{ij}, R_{in}, D_{ij}, and E_{in} change during learning; their dynamics are what is to be determined, from the dynamics of microscopic variables, i.e. connection weights.

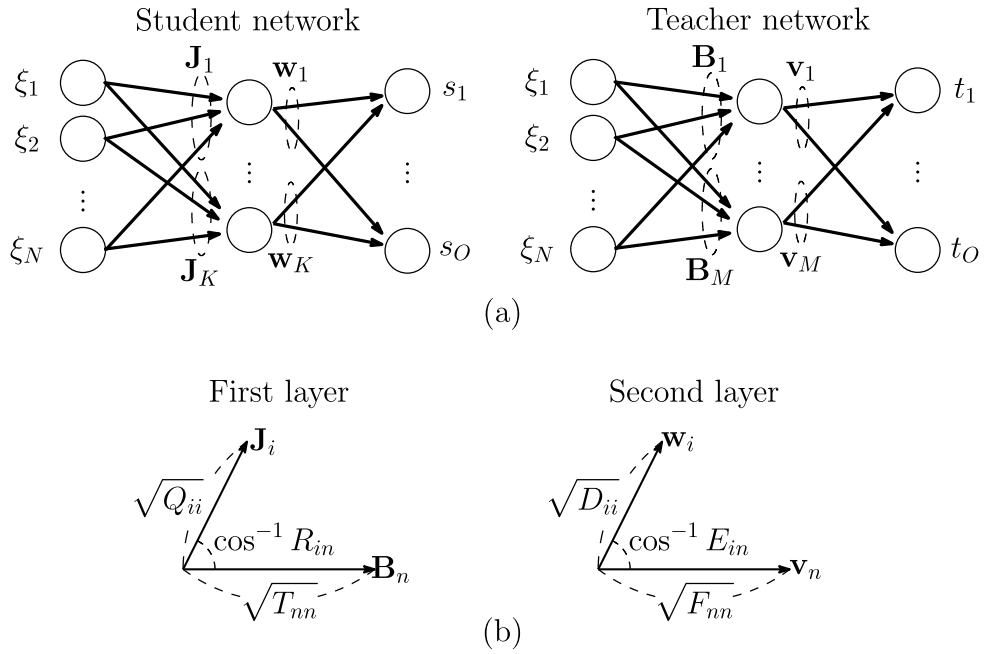

Figure 1. (a) Student and teacher networks. (b) Geometrical interpretation of order parameters Q_{ij}, R_{in}, T_{nm}, D_{ij}, E_{in}, and F_{nm}.

3. Dynamics of order parameters

In this paper, we adopt the stochastic gradient descent (SGD) learning algorithm, which underlies all conventional algorithms of neural networks used in practice. That is, every time an input sample is given, the output and sample loss is computed, the differentiation of the sample loss with respect to model parameters is computed, and finally, model parameters are moved against the gradient of the loss. The update rule of connection weights with SGD is written as

$$\begin{aligned}
\Delta \boldsymbol{J}_i &= -\frac{\eta}{N}\frac{\mathrm{d}\varepsilon}{\mathrm{d}\boldsymbol{J}_i} = \frac{\eta}{N}[(\boldsymbol{t}-\boldsymbol{s})\cdot\boldsymbol{w}_i]g'(x_i)\boldsymbol{\xi} \\
&= \frac{\eta}{N}\left[\left(\sum_{n=1}^{M}\boldsymbol{v}_n g(y_n) - \sum_{j=1}^{K}\boldsymbol{w}_j g(x_j)\right)\cdot\boldsymbol{w}_i\right]g'(x_i)\boldsymbol{\xi}, \qquad (1)\\
\Delta \boldsymbol{w}_i &= -\frac{\eta}{N}\frac{\mathrm{d}\varepsilon}{\mathrm{d}\boldsymbol{w}_i} = \frac{\eta}{N}g(x_i)(\boldsymbol{t}-\boldsymbol{s}) \\
&= \frac{\eta}{N}g(x_i)\left(\sum_{n=1}^{M}\boldsymbol{v}_n g(y_n) - \sum_{j=1}^{K}\boldsymbol{w}_j g(x_j)\right),
\end{aligned}$$

in which we set the learning rate as η/N, in order to obtain an N-independent macroscopic system. The first equation of (1) gives the update rule of order parameters Q_{ij} and R_{in} in the form of difference equations:

$$
\begin{aligned}
\Delta Q_{ij} &= (\boldsymbol{J}_i + \Delta\boldsymbol{J}_i)\cdot(\boldsymbol{J}_j + \Delta\boldsymbol{J}_j) - \boldsymbol{J}_i\cdot\boldsymbol{J}_j \\
&= \boldsymbol{J}_i\cdot\Delta\boldsymbol{J}_j + \boldsymbol{J}_j\cdot\Delta\boldsymbol{J}_i + \Delta\boldsymbol{J}_i\cdot\Delta\boldsymbol{J}_j \\
&= \frac{\eta}{N}\left[\sum_{p=1}^{M} E_{ip}g'(x_i)x_j g(y_p) - \sum_{p=1}^{K} D_{ip}g'(x_i)x_j g(x_p)\right. \\
&\quad \left. + \sum_{p=1}^{M} E_{jp}g'(x_j)x_i g(y_p) - \sum_{p=1}^{K} D_{jp}g'(x_j)x_i g(x_p)\right] \\
&\quad + \frac{\eta^2}{N^2}\|\boldsymbol{\xi}\|^2\left[\sum_{p,q}^{K,K} D_{ip}D_{jq}g'(x_i)g'(x_j)g(x_p)g(x_q)\right. \\
&\quad + \sum_{p,q}^{M,M} E_{ip}E_{jq}g'(x_i)g'(x_j)g(y_p)g(y_q) \\
&\quad \left. - \sum_{p,q}^{K,M} D_{ip}E_{jq}g'(x_i)g'(x_j)g(x_p)g(y_q) - \sum_{p,q}^{M,K} E_{ip}D_{jq}g'(x_i)g'(x_j)g(y_p)g(x_q)\right], \qquad (2) \\
\Delta R_{in} &= (\boldsymbol{J}_i + \Delta\boldsymbol{J}_i)\cdot\boldsymbol{B}_n - \boldsymbol{J}_i\cdot\boldsymbol{B}_n \\
&= \Delta\boldsymbol{J}_i\cdot\boldsymbol{B}_n \\
&= \frac{\eta}{N}\left[\sum_{p=1}^{M} E_{ip}g'(x_i)y_n g(y_p) - \sum_{p=1}^{K} D_{ip}g'(x_i)y_n g(x_p)\right].
\end{aligned}
$$

Since $\|\boldsymbol{\xi}\|^2 \approx N$ and the right hand sizes of these equations are $O(N^{-1})$, we can replace these difference equations with differential ones with $N \to \infty$, by taking the expectation over all input vectors $\boldsymbol{\xi}$:

$$
\begin{aligned}
\frac{\mathrm{d}Q_{ij}}{\mathrm{d}\tilde{\alpha}} &= \eta\left[\sum_{p=1}^{M} E_{ip}I_3(x_i,x_j,y_p) - \sum_{p=1}^{K} D_{ip}I_3(x_i,x_j,x_p)\right. \\
&\quad \left. + \sum_{p=1}^{M} E_{jp}I_3(x_j,x_i,y_p) - \sum_{p=1}^{K} D_{jp}I_3(x_j,x_i,x_p)\right] \\
&\quad + \eta^2\left[\sum_{p,q}^{K,K} D_{ip}D_{jq}I_4(x_i,x_j,x_p,x_q) + \sum_{p,q}^{M,M} E_{ip}E_{jq}I_4(x_i,x_j,y_p,y_q)\right. \\
&\quad \left. - \sum_{p,q}^{K,M} D_{ip}E_{jq}I_4(x_i,x_j,x_p,y_q) - \sum_{p,q}^{M,K} E_{ip}D_{jq}I_4(x_i,x_j,y_p,x_q)\right], \qquad (3) \\
\frac{\mathrm{d}R_{in}}{\mathrm{d}\tilde{\alpha}} &= \eta\left[\sum_{p=1}^{M} E_{ip}I_3(x_i,y_n,y_p) - \sum_{p=1}^{K} D_{ip}I_3(x_i,y_n,x_p)\right]
\end{aligned}
$$

$$
\begin{aligned}
&\text{where} \quad I_3(z_1,z_2,z_3) := \langle g'(z_1)z_2 g(z_3)\rangle, \\
&\qquad\quad I_4(z_1,z_2,z_3,z_4) := \langle g'(z_1)g'(z_2)g(z_3)g(z_4)\rangle.
\end{aligned} \qquad (4)
$$

In these equations, $\tilde{\alpha} := \alpha/N$ represents time (normalized number of steps), and the brackets $\langle\cdot\rangle$ represent the expectation when the input $\boldsymbol{\xi}$ follows $\mathcal{N}(\xi_i|0,1)$, that is, when

$(x_1, \ldots, x_K, y_1, \ldots, y_M)$ follows $\mathcal{N}(\mathbf{0}, \begin{pmatrix} Q & R \\ R^T & T \end{pmatrix})$. Likewise, the difference equations of the second layers' order parameters D_{ij} and E_{in} are obtained by the second equation of (1) as

$$\begin{aligned}
\Delta D_{ij} &= (\boldsymbol{w}_i + \Delta \boldsymbol{w}_i) \cdot (\boldsymbol{w}_j + \Delta \boldsymbol{w}_j) - \boldsymbol{w}_i \cdot \boldsymbol{w}_j \\
&= \boldsymbol{w}_i \cdot \Delta \boldsymbol{w}_j + \boldsymbol{w}_j \cdot \Delta \boldsymbol{w}_i + \Delta \boldsymbol{w}_i \cdot \Delta \boldsymbol{w}_j \\
&= \frac{\eta}{N} \left[\sum_{p=1}^{M} E_{ip} g(x_j) g(y_p) - \sum_{p=1}^{K} D_{ip} g(x_j) g(x_p) \right. \\
&\quad \left. + \sum_{p=1}^{M} E_{jp} g(x_i) g(y_p) - \sum_{p=1}^{K} D_{jp} g(x_i) g(x_p) \right] \\
&\quad + \frac{\eta^2}{N^2} \left[\sum_{p,q}^{K,K} D_{pq} g(x_i) g(x_j) g(x_p) g(x_q) + \sum_{p,q}^{M,M} F_{pq} g(x_i) g(x_j) g(y_p) g(y_q) \right. \\
&\quad \left. -2 \sum_{p,q}^{K,M} E_{pq} g(x_i) g(x_j) g(x_p) g(y_q) \right], \qquad (5) \\
\Delta E_{in} &= (\boldsymbol{w}_i + \Delta \boldsymbol{w}_i) \cdot \boldsymbol{v}_n - \boldsymbol{w}_i \cdot \boldsymbol{v}_n \\
&= \Delta \boldsymbol{w}_i \cdot \boldsymbol{v}_n \\
&= \frac{\eta}{N} \left[\sum_{p=1}^{M} F_{pn} g(x_i) g(y_p) - \sum_{p=1}^{K} E_{pn} g(x_i) g(x_p) \right].
\end{aligned}$$

Again, the right hand sides are $O(N^{-1})$, therefore these difference equations can be rewritten to differential versions with $N \to \infty$, by taking the expectation over all input vectors $\boldsymbol{\xi}$:

$$\begin{aligned}
\frac{\mathrm{d}D_{ij}}{\mathrm{d}\tilde{\alpha}} &= \eta \left[\sum_{p=1}^{M} E_{ip} I_2(x_j, y_p) - \sum_{p=1}^{K} D_{ip} I_2(x_j, x_p) \right. \\
&\quad \left. + \sum_{p=1}^{M} E_{jp} I_2(x_i, y_p) - \sum_{p=1}^{K} D_{jp} I_2(x_i, x_p) \right], \qquad (6) \\
\frac{\mathrm{d}E_{in}}{\mathrm{d}\tilde{\alpha}} &= \eta \left[\sum_{p=1}^{M} F_{pn} I_2(x_i, y_p) - \sum_{p=1}^{K} E_{pn} I_2(x_i, x_p) \right] \\
&\text{where} \quad I_2(z_1, z_2) := \langle g(z_1) g(z_2) \rangle. \qquad (7)
\end{aligned}$$

These differential equations govern the macroscopic dynamics. Note that the generalization loss ε_g, the expectation of loss value $\varepsilon(\boldsymbol{\xi}) = \frac{1}{2}\|\boldsymbol{s} - \boldsymbol{t}\|^2$ over all input vectors $\boldsymbol{\xi}$, is represented as

$$\begin{aligned}
\varepsilon_g &= \langle \frac{1}{2} \|\boldsymbol{s} - \boldsymbol{t}\|^2 \rangle \\
&= \frac{1}{2} \left[\sum_{p,q}^{M,M} F_{pq} I_2(y_p, y_q) + \sum_{p,q}^{K,K} D_{pq} I_2(x_p, x_q) - 2 \sum_{p,q}^{K,M} E_{pq} I_2(x_p, y_q) \right]. \qquad (8)
\end{aligned}$$

Expectation terms I_2, I_3 and I_4 can be analytically determined for some activation functions g, including sigmoid-like $g(x) = \mathrm{erf}(x/\sqrt{2})$ [4], and $g(x) = \mathrm{ReLU}(x) =: \max\{0, x\}$ which is commonly used in deep learning [22, 23].

4. Singular regions and plateaus

In this section we review the concept of singular regions as the cause of the plateau phenomenon [14].

In general, the input–output relationship (i.e. mapping) of a neural network is determined by the parameter values (i.e. connection weights) of the network. However, this correspondence is not one-to-one; in other words, there could be multiple settings of parameter values that result in one specific model (i.e. input–output mapping). For example, consider two-layer networks that have two hidden units (i.e. $K = 2$) and one output unit (i.e. $O = 1$) as

$$s = w_1 g(\boldsymbol{J}_1 \cdot \boldsymbol{\xi}) + w_2 g(\boldsymbol{J}_2 \cdot \boldsymbol{\xi}) \tag{9}$$

where the parameter is $(\boldsymbol{J}_1, \boldsymbol{J}_2, w_1, w_2)$. Then, all of the parameter settings in the parameter regions

$$\begin{aligned} \mathcal{R}_1(\boldsymbol{J}^*, w^*) &:= \{\boldsymbol{J}_1 = \boldsymbol{J}_2 = \boldsymbol{J}^* \text{ and } w_1 + w_2 = w^* \mid (\boldsymbol{J}_1, \boldsymbol{J}_2, w_1, w_2)\}, \\ \mathcal{R}_2(\boldsymbol{J}^*, w^*) &:= \{(\boldsymbol{J}_1 = \boldsymbol{J}^* \text{ and } w_1 = w^* \text{ and } w_2 = 0) \\ &\qquad \text{or } (\boldsymbol{J}_2 = \boldsymbol{J}^* \text{ and } w_1 = 0 \text{ and } w_2 = w^*) \mid (\boldsymbol{J}_1, \boldsymbol{J}_2, w_1, w_2)\} \end{aligned} \tag{10}$$

correspond to the same model (i.e. input–output mapping):

$$s = w^* g(\boldsymbol{J}^* \cdot \boldsymbol{\xi}). \tag{11}$$

These parameter regions $\mathcal{R}_i$ are called the singular region.

The model (11) can be regarded as a $K = 1$ model, parameterized by $\boldsymbol{J}^*$ and w^*. Now suppose that

$$\frac{\partial \varepsilon_g}{\partial w^*} = 0 \quad \text{and} \quad \frac{\partial \varepsilon_g}{\partial \boldsymbol{J}^*} = \boldsymbol{0}; \tag{12}$$

note that this occurs when the $K = 1$ model (11) gives a local minima of the generalization loss ε_g in the $K = 1$ parameter space. Then, one can show that the gradient of the generalization loss is also zero throughout the singular regions $\mathcal{R}_1$ and $\mathcal{R}_2$ in the original $K = 2$ parameter space; that is, $\partial \varepsilon_g / \partial w_i = 0$ and $\partial \varepsilon_g / \partial \boldsymbol{J}_i = \boldsymbol{0}$ if $(\boldsymbol{J}_1, \boldsymbol{J}_2, w_1, w_2) \in \mathcal{R}_1 \cup \mathcal{R}_2$, provided that (12) holds. Moreover, the singular region $\mathcal{R}_1$ has the following properties [24]:

- The region $\mathcal{R}_1$ is partially stable. When the parameter value is in the stable part of $\mathcal{R}_1$, it undergoes a long period of random walk, by fluctuations due to the random sampling of $\boldsymbol{\xi}$. Once the parameter value has reached the unstable part of the region, it starts moving away from the region.
- The region $\mathcal{R}_1$ is a Milnor-like attractor [19] in the sense that it has a positive measure of basin of attraction. This means that a randomly chosen initial parameter value will get trapped in the region with nonzero probability.

From these points, the singular region is completely different from a saddle point. When trapped in the singular region, the learning process inevitably slows down. That is why the singular region is considered to be the cause of plateaus.

However, the concept of problematic plateaus described above may not be true when a network has multiple outputs ($O > 1$); the loss gradient does not necessarily vanish at the

singular region, and the network might not be attracted to the singular region [21]. Thus, we analyzed the learning dynamics of networks that have multiple output units and examined whether or not the networks were trapped in the singular region and plateaus during learning.

5. Numerical results

We discuss the dynamics of learning in a two-layer perceptron by numerically solving the differential equations of the order parameters (3) and (6). For simplicity, we focus on the case with $K = M = 2$ units in the hidden layers and $O = 2$ units in the output layers. In the numerical experiments described below, we initialize the weights of the first layers of the student and teacher networks with $(\boldsymbol{J}_i)_j, (\boldsymbol{B}_i)_j \sim \mathcal{N}(0, 1/N)$ (i.i.d.). This initialization makes the initial values of the first layers' order parameters

$$Q = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad R = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad T = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{13}$$

on average. When N is finite, each component of the matrix Q, R and T has $O(N^{-1})$ noise; it vanishes as $N \to \infty$. It is critical how we initialize the weights of the second layers of the student and teacher networks, $\boldsymbol{v}_i$ and $\boldsymbol{w}_i$. For example, consider the case

$$\boldsymbol{w}_1 = w_1\boldsymbol{c}, \quad \boldsymbol{w}_2 = w_2\boldsymbol{c}; \quad \boldsymbol{v}_1 = v_1\boldsymbol{c}, \quad \boldsymbol{v}_2 = v_2\boldsymbol{c} \tag{14}$$

where $\boldsymbol{c}$ is an arbitrary two-dimensional constant vector whose norm is 1. In this setting, the outputs of the teacher and student networks, $\boldsymbol{s}, \boldsymbol{t} \in \mathbb{R}^2$, are confined in the one-dimensional subspace along $\boldsymbol{c}$. This makes the learning process equivalent to one with one output unit. In fact, when (14) holds,

$$\boldsymbol{t} = \left[\sum_{n=1}^{M} v_n g(\boldsymbol{B}_n \cdot \boldsymbol{\xi})\right] \boldsymbol{c} = t\boldsymbol{c}, \qquad \boldsymbol{s} = \left[\sum_{i=1}^{K} w_i g(\boldsymbol{J}_i \cdot \boldsymbol{\xi})\right] \boldsymbol{c} = s\boldsymbol{c} \tag{15}$$

where we defined scalar t and s as $\sum_{n=1}^{M} v_n g(\boldsymbol{B}_n \cdot \boldsymbol{\xi})$ and $\sum_{i=1}^{K} w_i g(\boldsymbol{J}_i \cdot \boldsymbol{\xi})$, respectively. And from the update rule (1),

$$\begin{aligned} \Delta\boldsymbol{J}_i &= \frac{\eta}{N}[(\boldsymbol{t} - \boldsymbol{s}) \cdot \boldsymbol{w}_i] g'(x_i)\boldsymbol{\xi} = \frac{\eta}{N} w_i (t - s) g'(x_i)\boldsymbol{\xi}, \\ \Delta\boldsymbol{w}_i &= \frac{\eta}{N} g(x_i)(\boldsymbol{t} - \boldsymbol{s}) = \frac{\eta}{N} g(x_i)(t - s)\boldsymbol{c} \\ &\text{that is,} \quad \Delta w_i = \frac{\eta}{N} g(x_i)(t - s) \end{aligned} \tag{16}$$

which is simply the update rule when both networks have only one output unit.

Thus, how the student network with two-dimensional output learns the teacher network with two-dimensional output largely depends on the initial condition of the weight matrices of their second layers. To see this, we consider an initial condition parameterized by θ:

$$\boldsymbol{w}_1 = \boldsymbol{c}, \quad \boldsymbol{w}_2 = \boldsymbol{c}_\theta; \quad \boldsymbol{v}_1 = \boldsymbol{c}, \quad \boldsymbol{v}_2 = \boldsymbol{c}_\theta \tag{17}$$

where $\boldsymbol{c}$ is again an arbitrary two-dimensional constant vector whose norm is 1, and $\boldsymbol{c}_\theta$ is a constant vector which is obtained by rotating $\boldsymbol{c}$ by θ. We refer to the parameter θ as non-degeneracy because it regulates the degeneracy of the weight matrices of the second layers of both networks. We can test various situations by changing θ continuously; $\theta = 0$ makes both matrices degenerate, and $\theta = \pi/2$ makes both matrices orthogonal. The former situation, $\theta = 0$, is

equivalent to the one-dimensional output situation, as previously described. The initial condition of the second layers' order parameters D_{ij}, E_{in}, F_{nm}, corresponding to (17), is given by

$$D = E = F = \begin{pmatrix} 1 & \cos\theta \\ \cos\theta & 1 \end{pmatrix}. \tag{18}$$

Putting these initial conditions together, we examined the learning dynamics in two ways: by simulating the original microscopic system with finite N, i.e. conducting stochastic gradient descent in accordance with the update rule (1), and by solving the differential equations (3) and (6) of the order parameters numerically under initial conditions that match the initial weights used when simulating the original microscopic system. The black lines in figure 2 show the time courses of the generalization loss ε_g in several typical situations: (a) $\theta = 0$, (b) $\theta = \pi/8$, (c) $\theta = \pi/4$, and (d) $\theta = \pi/2$. To evaluate quantitatively how near the student network is to the singular region $\mathcal{R}_1$ and $\mathcal{R}_2$, we calculated two measures: the overlap of vectors $\boldsymbol{J}_1$ and $\boldsymbol{J}_2$, i.e. $m_{12}^{(1)} := |\boldsymbol{J}_1 \cdot \boldsymbol{J}_2|/\|\boldsymbol{J}_1\|\|\boldsymbol{J}_2\| = |Q_{12}|/\sqrt{Q_{11}Q_{22}}$, and the minimum norm of vectors $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$, i.e. $l_{\min}^{(2)} := \min\{\|\boldsymbol{w}_1\|, \|\boldsymbol{w}_2\|\} = \min\{\sqrt{D_{11}}, \sqrt{D_{22}}\}$. Note that $m_{12}^{(1)}$ measures proximity to the region $\mathcal{R}_1$, and $l_{\min}^{(2)}$ indicates the distance to the region $\mathcal{R}_2$. Figure 2 also shows the time evolutions of these measures with blue and red lines, respectively. Results of microscopic simulations are also shown by dots. In every plot in figure 2, the solid lines and dots agree well, meaning that the macroscopic system of order parameters appropriately captures the microscopic system of connection weights. In figure 2(a), the generalization loss ε_g stops during the first $\sim$1800 steps, along with high values of $m_{12}^{(1)}$, meaning that falling into the singular region $\mathcal{R}_1$ is derived from the network's symmetry. In figure 2(b), the plateau shortens. An increase in $m_{12}^{(1)}$ is still observed, although its peak is lower than figure 2(a). In figures 2(c) and (d), there is no apparent plateau. In particular, the overlap $m_{12}^{(1)}$ is always 0 in figure 2(d), signifying that the student network does not approach the singular region $\mathcal{R}_1$ at all during learning. Note also that no approach to the singular region $\mathcal{R}_2$, indicated by high values of $l_{\min}^{(2)}$, is observed in any plot in figure 2 at any time. Figure 3 shows the times where the plateau begins and ends, depending on θ, calculated by the numerical solutions of the order parameters. Here we define plateaus as 'where the logarithm of generalization loss decreases at a rate slower than a half of the typical rate', where the typical rate is measured as the rate when $\varepsilon_g < 10^{-10}$ is achieved. We found that the plateau is observed if and only if roughly $|\theta| < 0.1\pi$ holds. Note that our quantitative definition of plateaus above contains some arbitrariness, but it does not affect the main point; the plateau phenomenon get alleviated as $|\theta|$ increases, as is evident from figure 2.

6. Cases for orthogonal second layers

Up to this point, we considered cases in which the second layers of two networks have identical weights, as (17). However, this is not practical because it means that the student knows about the teacher in advance. Thus, we consider a slightly different situation:

$$\boldsymbol{w}_1 = \boldsymbol{c}, \quad \boldsymbol{w}_2 = \boldsymbol{c}_{\pi/2}; \quad \boldsymbol{v}_1 = \boldsymbol{c}_{\phi}, \quad \boldsymbol{v}_2 = \boldsymbol{c}_{\phi+\pi/2}, \tag{19}$$

wherein the student matrix and teacher matrix are not identical but are both orthogonal. The initial conditions of the order parameters D_{ij}, E_{in}, F_{nm}, corresponding to (19), are given by

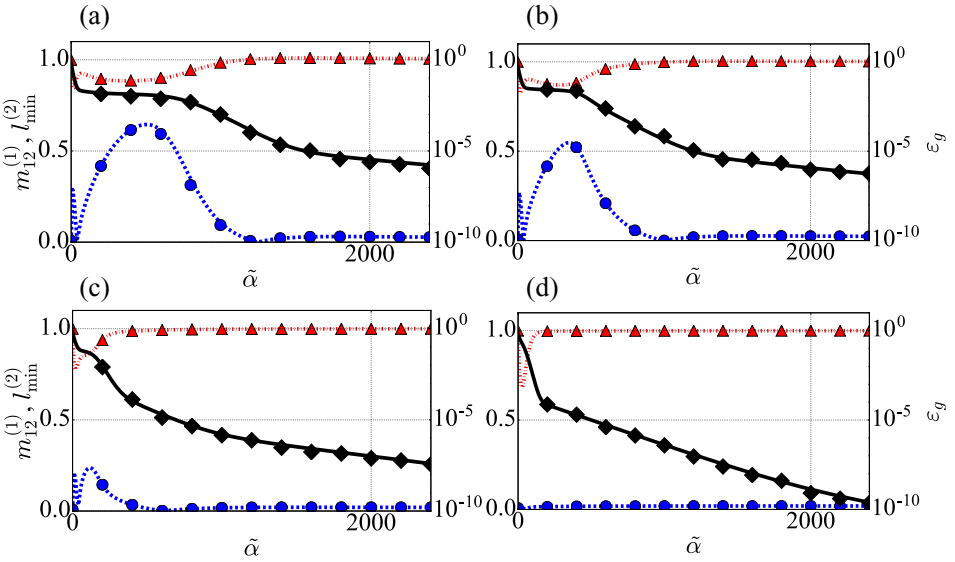

Figure 2. Dependence of time course of generalization loss ε_g (black solid line) on non-degeneracy parameter θ. Time course of student's first layer's overlap $m_{12}^{(1)} := |Q_{12}|/\sqrt{Q_{11}Q_{22}}$ (blue dashed line) and minimum norm of student's second norm $l_{\min}^{(2)} := \min\{\sqrt{D_{11}}, \sqrt{D_{22}}\}$ (red dot-dashed line) also shown. These lines indicate how student network is close to singular regions $\mathcal{R}_1$ and $\mathcal{R}_2$, respectively. (a) $\theta = 0$, (b) $\theta = \pi/8$, (c) $\theta = \pi/4$, (d) $\theta = \pi/2$. Simulation results of microscopic systems shown by dots (diamonds, circles, and triangles for ε_g, $m_{12}^{(1)}$ and $l_{\min}^{(2)}$, respectively). Generalization loss of simulation results approximated by averaged sample loss (ε) over 100 contiguous steps. Simulation parameters: $N = 1000$, $\eta = 0.1$ ($\eta/N = 0.0001$). Activation function: $g(x) = \mathrm{erf}(x/\sqrt{2})$.

$$D = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad E = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}, \quad F = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{20}$$

Under these initial conditions, we found that the orthogonality $\boldsymbol{w}_1 \perp \boldsymbol{w}_2$ remains true during learning at any time, as formally stated in the proposition below. In other words, we prove that the student network stays opposite the singular region $\mathcal{R}_1$.

Proposition. *Assume $K = M = 2$ and $O \geqslant 2$, and consider the situation in which $N \to \infty$. Assume that the activation g is an odd function and that the right sides of the differential equations are Lipschitz continuous in the vicinity of the solution trajectory (Q, R, D, E) during learning. If both the teacher and student network have a pair of orthogonal column vectors in its second layer's weight matrix, that is, $\boldsymbol{v}_1 \perp \boldsymbol{v}_2$ and $\boldsymbol{w}_1 \perp \boldsymbol{w}_2$ hold at the initial state, then the orthogonality $\boldsymbol{w}_1 \perp \boldsymbol{w}_2$ holds at any time during learning.* ■

We give the proof of the proposition in the appendix.

The numerical solution of the differential equations of the order parameters, under the initial conditions described above, is shown in figure 4. These solutions tell us that there is no approach to the singular regions $\mathcal{R}_1$ and $\mathcal{R}_2$ and that no plateau is seen, except for the case of

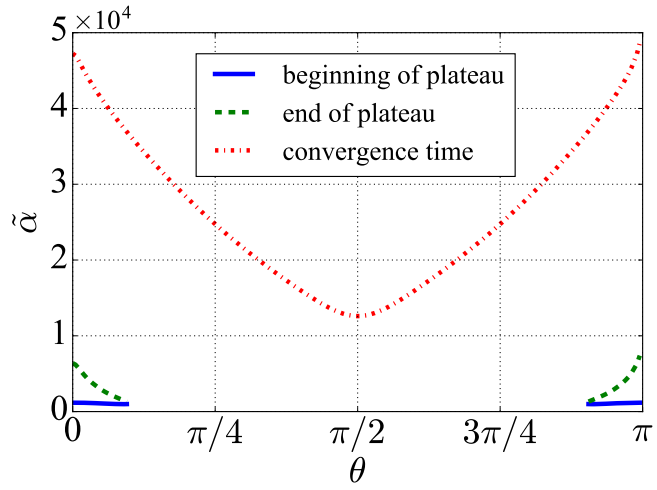

Figure 3. Dependence of when the plateau starts and ends and when generalization loss converges on non-degeneracy parameter θ. Blue solid line: start of the plateau, green dashed line: end of the plateau, red dot-dashed line: achieving $\varepsilon_g < 10^{-10}$. See main text for definition of the plateau in this figure. Parameters: $\eta = 0.02$. Activation: $g(x) = \mathrm{erf}(x/\sqrt{2})$.

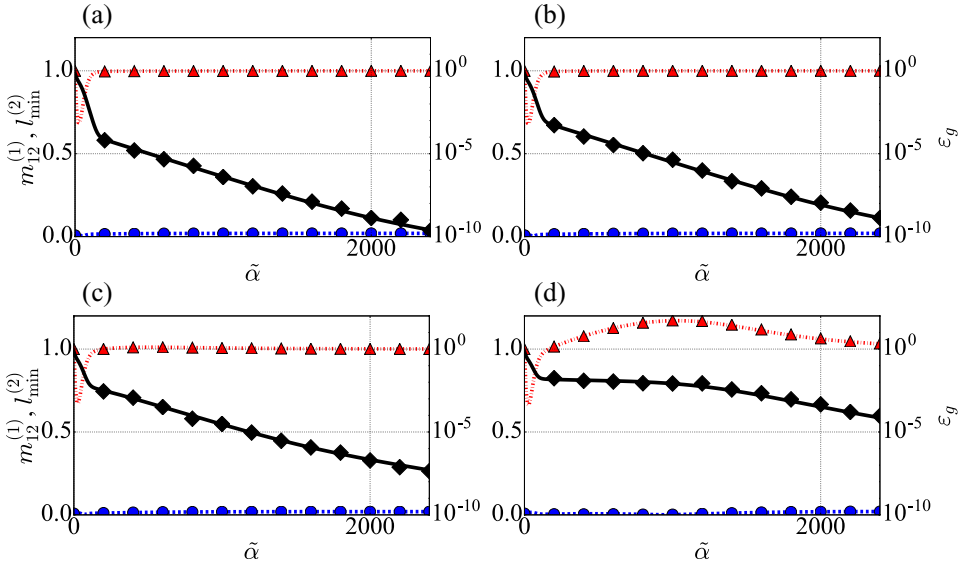

Figure 4. Dependence of time course of generalization loss ε_g (black solid line) on rotation parameter ϕ. $m_{12}^{(1)}$ (blue dashed line) and $l_{\min}^{(2)}$ (red dot-dashed line) also shown; see caption of figure 2 for a detailed explanation. (a) $\phi = 0$, (b) $\phi = \pi/8$, (c) $\phi = 3\pi/16$, (d) $\phi = \pi/4$. Simulation results of microscopic systems shown by dots (diamonds, circles, and triangles for ε_g, $m_{12}^{(1)}$ and $l_{\min}^{(2)}$, respectively). Generalization loss of simulation results approximated by averaged sample loss (ε) over 100 contiguous steps. Simulation parameters: $N = 1000$, $\eta = 0.1$ ($\eta/N = 0.0001$). Activation: $g(x) = \mathrm{erf}(x/\sqrt{2})$.

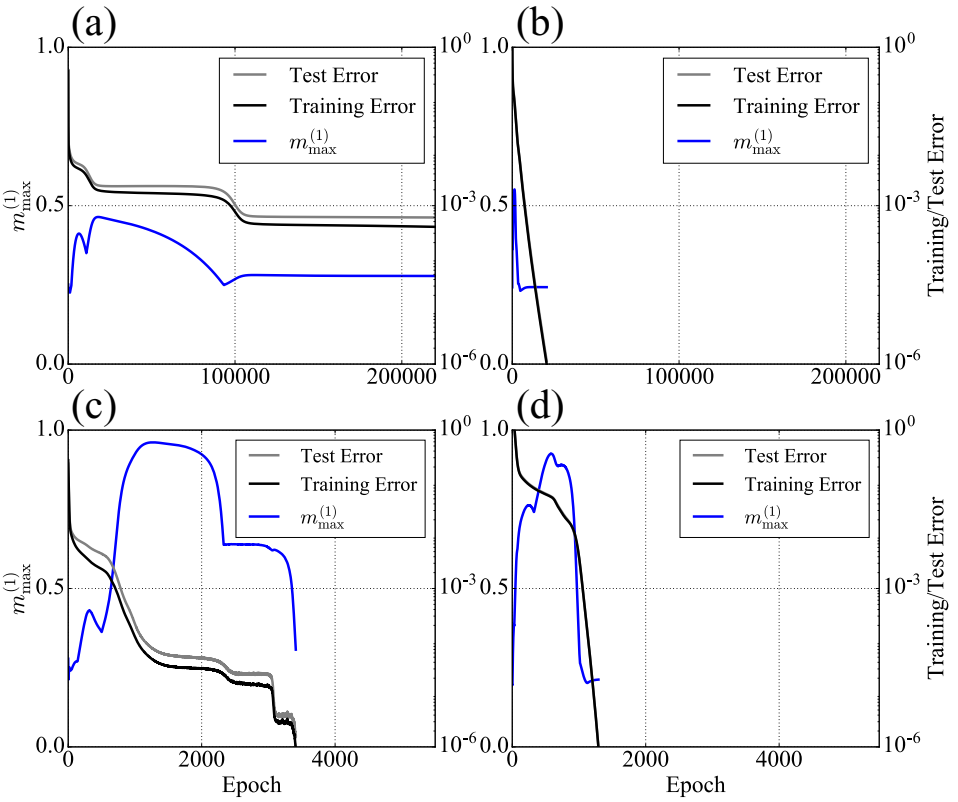

Figure 5. Simulation results of time course of training loss (black line) and test loss (gray line), depending on the number of output units and choice of optimizers. Time course of student's first layer's maximum overlap $m^{(1)}_{\max} := \max_{i<j} |Q_{ij}|/\sqrt{Q_{ii}Q_{jj}}$ (blue line) also shown. This maximum overlap indicates how the student network is close to the singular region. (a) and (b) Stochastic gradient descent (learning rate: 0.01), (c) and (d) Adam optimizer. (a) and (c) Networks with one output unit, (b) and (d) networks with ten output units. Mini-batch size: 1000. Activation function: $g(x) = \tanh(x)$.

$\phi = \pi/4$ (figure 4(d)); this plateau is not due to the singular regions but rather to being stuck at a saddle point where the weight vectors of the second layer of the student network are perturbed by the stochastic gradient and can easily escape.

7. Simulation results for more practical cases

By using order parameters we analyzed theoretically the case with two hidden units. However, networks with a greater number of hidden units are usually used in practice. Also, the setting above assumes an infinite number of samples which is not available in reality. Furthermore, various optimization techniques developed in recent years such as mini-batch, dropout and gradient descent with adaptive learning rate are widely used. In these practical cases the learning dynamics might not be tractable. We examined such cases by numerical simulations of microscopic systems.

In our experiment, a student network and a teacher network both of which have 100 input units and ten hidden units are used. First we generated 4000 training samples and 1000 test samples by using the teacher network whose weights are randomly chosen. We then trained the student network whose weights are randomly initialized using training samples only, and computed the training loss and the test loss after every epoch.

In figure 5, the dynamics of the learning in the experiment is shown. (Although this dynamics depends on the initial weights, the qualitative shapes of the plateaus do not; see supplemental material.) The output dimension is 1 in figures 5(a) and (c), and 10 in figures 5(b) and (d). These results apparently indicate that the plateau is alleviated by multiple output units.

We also examined the cases with an Adam optimizer, bias terms and dropout regularization. All these results are consistent with the idea that multiple outputs mitigate the plateaus due to the singular regions (see the supplemental material).

8. Conclusion

We analyzed the learning dynamics of two-layer networks that have multiple output units by means of statistical mechanical formulation. By defining order parameters, deriving the differential equations they follow, and solving said equations, we clarified experimentally and theoretically that multiple-output networks are less likely to be trapped in plateaus because of singularity than single-output networks are.

Through this paper, we suggest reconsidering the established idea that singular structures cause plateaus and they interrupt learning. However, more general cases, such as cases with deeper neural networks, have yet to be researched.

Acknowledgment

This work was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research(A) (No. 18H04106).

Appendix. Proof of continued orthogonality during learning

This section gives the proof of the proposition in the main text:

Proposition. *Assume $K = M = 2$ and $O \geqslant 2$, and consider the situation in which $N \to \infty$. Assume that the activation g is an odd function and that the right sides of the differential equations are Lipschitz continuous in the vicinity of the solution trajectory (Q, R, D, E) during learning. If both the teacher and student network have a pair of orthogonal column vectors in its second layer's weight matrix, that is, $\boldsymbol{v}_1 \perp \boldsymbol{v}_2$ and $\boldsymbol{w}_1 \perp \boldsymbol{w}_2$ hold at the initial state, then the orthogonality $\boldsymbol{w}_1 \perp \boldsymbol{w}_2$ holds at any time during learning.* ■

Proof. We prove the following claim: suppose $K = M = 2$ and $O \geqslant 2$, and the activation g is an odd function, then the differential equations (3) and (6) imply that

$$\begin{aligned} &\text{if} \quad Q \propto I, \quad R + R^T \propto I, \quad D \propto I, \quad E + E^T \propto I, \quad T \propto I, \quad F \propto I \quad (*) \\ &\text{then} \quad \dot{Q} \propto I, \quad \dot{R} + \dot{R}^T \propto I, \quad \dot{D} \propto I, \quad \dot{E} + \dot{E}^T \propto I, \end{aligned} \tag{A.1}$$

where I denotes the 2×2 identity matrix, and $X \propto I$ means that there exists $c \in \mathbb{R}$ such that $X = cI$.

Proving this claim (A.1) is sufficient for the proof of the proposition. Suppose that $\hat{\Theta}(\alpha) = (\hat{Q}(\alpha), \hat{R}(\alpha), \hat{D}(\alpha), \hat{E}(\alpha))$ is one solution of the differential equation. What we have to show is that the solution always lies in the subspace

$$S := \{(Q, R, D, E) \mid Q \propto I,\ R + R^T \propto I,\ D \propto I \text{ and } E + E^T \propto I\}. \tag{A.2}$$

If we denote by $P\hat{\Theta}$ the vector from $\hat{\Theta}$ to the foot of its perpendicular to S (note that this mapping is linear), what we have to show is

$$\boldsymbol{f}(\alpha) := P\hat{\Theta}(\alpha) = \mathbf{0} \tag{A.3}$$

for all time α. We have

$$\boldsymbol{f}(0) = \mathbf{0} \tag{A.4}$$

by substituting the initial condition (13) and (20), and we have

$$\begin{aligned} \boldsymbol{f}(\alpha) = \mathbf{0} \quad &\Longrightarrow \quad P\frac{\mathrm{d}}{\mathrm{d}\alpha}\hat{\Theta}(\alpha) = \mathbf{0} \quad \text{(from claim (A.1))} \\ &\Longrightarrow \quad \frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}\alpha}(\alpha) = P\frac{\mathrm{d}}{\mathrm{d}\alpha}\hat{\Theta}(\alpha) = \mathbf{0} \end{aligned} \tag{A.5}$$

for given time α. These imply that $\boldsymbol{f}(\alpha) \equiv \mathbf{0}$ is one solution of the differential equation for $\boldsymbol{f}$. Lipschitz continuity ensures that the uniqueness of the solution of the differential equation, therefore we have $\boldsymbol{f}(\alpha) \equiv \mathbf{0}$ for all α.

To prove the claim (A.1), we first show the following lemma. ■

Lemma A.1. *If the condition (∗) in the claim (A.1) holds, the following relations hold:*

$$\begin{aligned} I_2(a_{i_1}, b_{i_2}) &= (-1)^{\delta(i_1=1)\delta(i_2=1)} I_2(a_{j_1}, b_{j_2}), \\ I_3(a_{i_1}, b_{i_2}, c_{i_3}) &= (-1)^{\delta(i_2=1)\delta(i_3=1)} I_3(a_{j_1}, b_{j_2}, c_{j_3}), \\ I_4(a_{i_1}, b_{i_2}, c_{i_3}, d_{i_4}) &= (-1)^{\delta(i_3=1)\delta(i_4=1)} I_4(a_{j_1}, b_{j_2}, c_{j_3}, d_{j_4}), \end{aligned} \tag{A.6}$$

where each of a, b, c and d is either x or y, and $\{i_e, j_e\} = \{1, 2\}$ *for* $e = 1, 2, 3, 4$.

Proof of lemma 1. If the condition $(*)$ holds, the covariance matrix of (x_1, x_2, y_1, y_2) is given by

$$\Sigma = \begin{pmatrix} Q_{11} & 0 & R_{11} & R_{12} \\ 0 & Q_{11} & -R_{12} & R_{11} \\ R_{11} & -R_{12} & T_{11} & 0 \\ R_{12} & R_{11} & 0 & T_{11} \end{pmatrix}. \tag{A.7}$$

This matrix Σ has the following property:

$$\mathcal{N}(z_1, z_2, z_3, z_4 \mid 0, \Sigma) = \mathcal{N}(-z_2, z_1, -z_4, z_3 \mid 0, \Sigma). \tag{A.8}$$

Therefore, for arbitrary functions f,

$$\begin{aligned}\langle f(x_1,x_2,y_1,y_2)\rangle &= \int f(x_1,x_2,y_1,y_2)\mathcal{N}(x_1,x_2,y_1,y_2 \mid 0,\Sigma)\,\mathrm{d}x_1\mathrm{d}x_2\mathrm{d}y_1\mathrm{d}y_2\\ &= \int f(-x_2,x_1,-y_2,y_1)\mathcal{N}(-x_2,x_1,-y_2,y_1 \mid 0,\Sigma)\,\mathrm{d}x_1\mathrm{d}x_2\mathrm{d}y_1\mathrm{d}y_2\\ &= \int f(-x_2,x_1,-y_2,y_1)\mathcal{N}(x_1,x_2,y_1,y_2 \mid 0,\Sigma)\,\mathrm{d}x_1\mathrm{d}x_2\mathrm{d}y_1\mathrm{d}y_2\\ &= \langle f(-x_2,x_1,-y_2,y_1)\rangle,\end{aligned} \tag{A.9}$$

and since g is an odd function,

$$\begin{aligned}I_2(a_{i_1},b_{i_2}) &= \langle g(a_{i_1})g(b_{i_2})\rangle\\ &= \langle g((-1)^{\delta(i_1=1)}a_{j_1})g((-1)^{\delta(i_2=1)}b_{j_2})\rangle\\ &= (-1)^{\delta(i_1=1)\delta(i_2=1)}\langle g(a_{j_1})g(b_{j_2})\rangle\\ &= (-1)^{\delta(i_1=1)\delta(i_2=1)}I_2(a_{j_1},b_{j_2}),\\ I_3(a_{i_1},b_{i_2},c_{i_3}) &= \langle g'(a_{i_1})b_{i_2}g(c_{i_3})\rangle\\ &= \langle g'((-1)^{\delta(i_1=1)}a_{i_1})(-1)^{\delta(i_2=1)}b_{i_2}g((-1)^{\delta(i_3=1)}c_{i_3})\rangle\\ &= (-1)^{\delta(i_2=1)\delta(i_3=1)}\langle g'(a_{j_1})g(b_{j_2})\rangle\\ &= (-1)^{\delta(i_2=1)\delta(i_3=1)}I_3(a_{j_1},b_{j_2},c_{j_3}),\\ I_4(a_{i_1},b_{i_2},c_{i_3},d_{i_4}) &= \langle g'(a_{i_1})g'(b_{i_2})g(c_{i_3})g(d_{i_4})\rangle\\ &= \langle g'((-1)^{\delta(i_1=1)}a_{j_1})g'((-1)^{\delta(i_2=1)}b_{j_2})\\ &\qquad \cdot g((-1)^{\delta(i_3=1)}c_{j_3})g((-1)^{\delta(i_4=1)}d_{j_4})\rangle\\ &= (-1)^{\delta(i_3=1)\delta(i_4=1)}\langle g'(a_{j_1})g'(b_{j_2})g(c_{j_3})g(d_{j_4})\rangle\\ &= (-1)^{\delta(i_3=1)\delta(i_4=1)}I_4(a_{j_1},b_{j_2},c_{j_3},d_{j_4}),\end{aligned} \tag{A.10}$$

which is the statement of the lemma. ■

Lemma A.2. $I_4(z_1,z_2,z_3,z_4) = I_4(z_2,z_1,z_3,z_4) = I_4(z_1,z_2,z_4,z_3)$.

Proof of lemma 2. The proof is clear from the definition of I_4. ■

Proof of proposition. Since the matrices Q, T, D, and F are symmetric, what we have to show is

$$\begin{aligned}&\dot{Q}_{11} = \dot{Q}_{22}, \quad \dot{Q}_{12} = 0,\\ &\dot{R}_{11} = \dot{R}_{22}, \quad \dot{R}_{12} + \dot{R}_{21} = 0,\\ &\dot{D}_{11} = \dot{D}_{22}, \quad \dot{D}_{12} = 0,\\ &\dot{E}_{11} = \dot{E}_{22}, \quad \dot{E}_{12} + \dot{E}_{21} = 0.\end{aligned} \tag{A.11}$$

First, for $\dot{Q}$, we have

$$\begin{aligned}
N\dot{Q}_{11} &= 2\eta\left[E_{11}I_3(x_1,x_1,y_1) + E_{12}I_3(x_1,x_1,y_2) - D_{11}I_3(x_1,x_1,x_1)\right] \\
&\quad + \eta^2\left[D_{11}^2 I_4(x_1,x_1,x_1,x_1) + E_{11}^2 I_4(x_1,x_1,y_1,y_1)\right. \\
&\quad + E_{11}E_{12}[I_4(x_1,x_1,y_1,y_2) + I_4(x_2,x_2,y_2,y_1)] + E_{12}^2 I_4(x_1,x_1,y_2,y_2) \\
&\quad - D_{11}E_{11}[I_4(x_1,x_1,x_1,y_1) + I_4(x_1,x_1,y_1,x_1)] \\
&\quad \left. - D_{11}E_{12}[I_4(x_1,x_1,x_1,y_2) + I_4(x_1,x_1,y_2,x_1)]\right], \\
N\dot{Q}_{22} &= 2\eta\left[E_{11}I_3(x_2,x_2,y_2) - E_{12}I_3(x_2,x_2,y_1) - D_{11}I_3(x_2,x_2,x_2)\right] \\
&\quad + \eta^2\left[D_{11}^2 I_4(x_2,x_2,x_2,x_2) + E_{11}^2 I_4(x_2,x_2,y_2,y_2)\right. \\
&\quad - E_{11}E_{12}[I_4(x_2,x_2,y_2,y_1) + I_4(x_1,x_1,y_1,y_2)] + E_{12}^2 I_4(x_2,x_2,y_1,y_1) \\
&\quad - D_{11}E_{11}[I_4(x_2,x_2,x_2,y_2) + I_4(x_2,x_2,y_2,x_2)] \\
&\quad \left. + D_{11}E_{12}[I_4(x_2,x_2,x_2,y_1) + I_4(x_2,x_2,y_1,x_2)]\right], \\
N\dot{Q}_{12} &= \eta\left[E_{11}[I_3(x_1,x_2,y_1) + I_3(x_2,x_1,y_2)] + E_{12}[I_3(x_1,x_2,y_2) - I_3(x_2,x_1,y_1)]\right. \\
&\quad \left. - D_{11}[I_3(x_1,x_2,x_1) + I_3(x_2,x_1,x_2)]\right] \\
&\quad + \eta^2\left[D_{11}^2 I_4(x_1,x_2,x_1,x_2) + E_{11}^2 I_4(x_1,x_2,y_1,y_2)\right. \\
&\quad + E_{11}E_{12}[-I_4(x_1,x_2,y_1,y_1) + I_4(x_1,x_2,y_2,y_2)] - E_{12}^2 I_4(x_1,x_2,y_2,y_1) \\
&\quad - D_{11}E_{11}[I_4(x_1,x_2,x_1,y_2) + I_4(x_1,x_2,y_1,x_2)] \\
&\quad \left. - D_{11}E_{12}[-I_4(x_1,x_2,x_1,y_1) + I_4(x_1,x_2,y_2,x_2)]\right]
\end{aligned} \tag{A.12}$$

from the equation (3) and the assumption of the proposition. Applying the lemmas to each I-term, we can confirm $\dot{Q}_{11} = \dot{Q}_{22}$ and $\dot{Q}_{12} = 0$.

With respect to R, we find

$$\begin{aligned}
N\dot{R}_{11} &= \eta\left[E_{11}I_3(x_1,y_1,y_1) + E_{12}I_3(x_1,y_1,y_2) - D_{11}I_3(x_1,y_1,x_1)\right], \\
N\dot{R}_{22} &= \eta\left[E_{11}I_3(x_2,y_2,y_2) - E_{12}I_3(x_2,y_2,y_1) - D_{11}I_3(x_2,y_2,x_2)\right], \\
N\dot{R}_{12} &= \eta\left[E_{11}I_3(x_1,y_2,y_1) + E_{12}I_3(x_1,y_2,y_2) - D_{11}I_3(x_1,y_2,x_1)\right], \\
N\dot{R}_{21} &= \eta\left[E_{11}I_3(x_2,y_1,y_2) - E_{12}I_3(x_2,y_1,y_1) - D_{11}I_3(x_2,y_1,x_2)\right],
\end{aligned} \tag{A.13}$$

and by applying the lemmas to the I-terms, $\dot{R}_{11} = \dot{R}_{22}$ and $\dot{R}_{12} + \dot{R}_{21} = 0$ are implied.

In the same way, we have

$$\begin{aligned}
N\dot{D}_{11} &= 2\eta\left[E_{11}I_2(x_1,y_1) + E_{12}I_2(x_1,y_2) - D_{11}I_2(x_1,x_1)\right], \\
N\dot{D}_{22} &= 2\eta\left[E_{11}I_2(x_2,y_2) - E_{12}I_2(x_2,y_1) - D_{11}I_2(x_2,x_2)\right], \\
N\dot{D}_{12} &= \eta[E_{11}[I_2(x_2,y_1) + I_2(x_1,y_2)] + E_{12}[I_2(x_2,y_2) - I_2(x_1,y_1)] \\
&\quad - D_{11}[I_2(x_2,x_1) + I_2(x_1,x_2)]],
\end{aligned} \tag{A.14}$$

and

$$\begin{aligned}
N\dot{E}_{11} &= \eta\left[F_{11}I_2(x_1,y_1) - E_{11}I_2(x_1,x_1) + E_{12}I_2(x_1,x_2)\right], \\
N\dot{E}_{22} &= \eta\left[F_{11}I_2(x_2,y_2) - E_{11}I_2(x_2,x_2) - E_{12}I_2(x_2,x_1)\right], \\
N\dot{E}_{12} &= \eta\left[F_{11}I_2(x_1,y_2) - E_{11}I_2(x_1,x_2) - E_{12}I_2(x_1,x_1)\right], \\
N\dot{E}_{21} &= \eta\left[F_{11}I_2(x_2,y_1) - E_{11}I_2(x_2,x_1) + E_{12}I_2(x_2,x_2)\right],
\end{aligned} \tag{A.15}$$

and by using the lemmas we can confirm $\dot{D}_{11} = \dot{D}_{22}$, $\dot{D}_{12} = 0$, $\dot{E}_{11} = \dot{E}_{22}$, and $\dot{E}_{12} + \dot{E}_{21} = 0$, the statements of the proposition. ■

ORCID iDs

Yuki Yoshida https://orcid.org/0000-0002-1402-7840

References

[1] Rosenblatt F 1958 *Psychol. Rev.* **65** 386
[2] Rumelhart D E, Hinton G E and Williams R J 1986 *Nature* **323** 533–6
[3] Riegler P and Biehl M 1995 *J. Phys. A: Math. Gen.* **28** L507
[4] Saad D and Solla S A 1995 *Phys. Rev.* E **52** 4225
[5] Biehl M, Riegler P and Wöhler C 1996 *J. Phys. A: Math. Gen.* **29** 4769
[6] Freeman J A and Saad D 1997 *Neural Comput.* **9** 1601–22
[7] Huh N, Oh J and Kang K 2000 *J. Phys. A: Math. Gen.* **33** 8663
[8] Park H, Amari S and Fukumizu K 2000 *Neural Netw.* **13** 755–64
[9] Inoue M, Park H and Okada M 2003 *J. Phys. Soc. Japan* **72** 805–10
[10] Park H, Inoue M and Okada M 2005 *Prog. Theor. Phys. Suppl.* **157** 275–9
[11] Saxe A M, McClelland J L and Ganguli S 2014 *Int. Conf. on Learning Representations* (arXiv: 1312.6120) accepted
[12] Advani M S and Saxe A M 2017 (arXiv:1710.03667)
[13] Saxe A M, McClelland J L and Ganguli S 2018 (arXiv:1810.10531)
[14] Fukumizu K and Amari S 2000 *Neural Netw.* **13** 317–27
[15] Amari S and Ozeki T 2001 *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.* **84** 31–8
[16] Cousseau F, Ozeki T and Amari S 2008 *IEEE Trans. Neural Netw.* **19** 1313–28
[17] Amari S, Park H and Ozeki T 2006 *Neural Comput.* **18** 1007–65
[18] Amari S and Nagaoka H 2000 *Methods of Information Geometry* (*Translations of Mathematical Monographs* vol 191) (Providence, RI: AMS) (https://doi.org/10.1090/mmono/191)
[19] Milnor J 1985 *The Theory of Chaotic Attractors* (Berlin: Springer) pp 243–64
[20] Biehl M and Schwarze H 1995 *J. Phys. A: Math. Gen.* **28** 643
[21] Amari S, Ozeki T, Karakida R, Yoshida Y and Okada M 2018 *Neural Comput.* **30** 1–33
[22] LeCun Y, Bengio Y and Hinton G 2015 *Nature* **521** 436–44
[23] Yoshida Y, Karakida R, Okada M and Amari S 2017 *J. Phys. Soc. Japan* **86** 044002
[24] Wei H, Zhang J, Cousseau F, Ozeki T and Amari S 2008 *Neural Comput.* **20** 813–43