Centre manifold analysis of plateau phenomena in learning of three-layer perceptron

Daiji Tsutsui
Department of Mathematics, Osaka University
Toyonaka, Osaka 560-0043, Japan
E-mail: d-tsutsui@cr.math.sci.osaka-u.ac.jp

Abstract

A three-layer perceptron is the most basic model of hierarchical neural networks. We treat a gradient system representing the learning process of the three-layer perceptron. In its parameter space, a three-layer perceptron has one-dimensional singular regions comprising both attractive and repulsive parts, which is often called a Milnor-like attractor. In this paper, we introduce an analysis of the learning process in the vicinity of a Milnor-like attractor based on the centre manifold theory.

This paper is related to the article [3], which is published in *Neural Computation*.

1 Backgrounds

1.1 Gradient descent method

Mathematically, a three-layer perceptron is a family of functions given by

$$\begin{aligned} \boldsymbol{f}_{(d)}(\boldsymbol{x};\boldsymbol{\theta}) &= \sum_{i=1}^{d} \boldsymbol{v}_i\, \varphi\left(\boldsymbol{w}_i \cdot \boldsymbol{x} + b_i\right), \quad \boldsymbol{x} \in \mathbb{R}^n, \\ \boldsymbol{\theta} &= (\boldsymbol{w}_1, \dots, \boldsymbol{w}_d, b_1, \dots, b_d, \boldsymbol{v}_1, \dots, \boldsymbol{v}_d), \end{aligned} \tag{1}$$

where $\boldsymbol{\theta}$ is a system parameter with $\boldsymbol{w}_1, \cdots, \boldsymbol{w}_d \in \mathbb{R}^n$ being the weight vectors for the second layer, $b_1, \dots, b_d \in \mathbb{R}$ the bias terms for the second, $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_d \in \mathbb{R}^m$ the

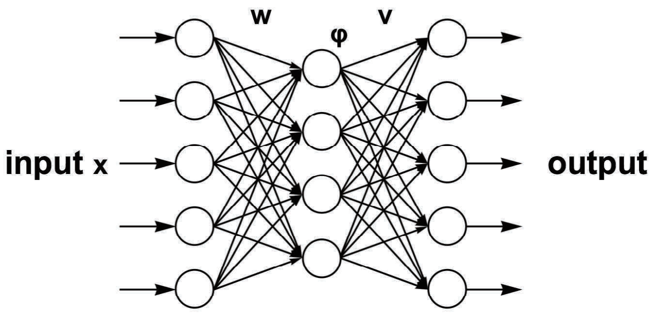


Figure 1: A schematic diagram of a three-layer perceptron presented in (2).

weight vectors for the third, and $\varphi : \mathbb{R} \to \mathbb{R}$ is an activation function. Throughout this paper, we assume that the activation function φ is twice differentiable. We shall call the function (1) an (n-d-m)-perceptron. The numbers n and m are fixed at the outset as the sizes of input and output vectors, while the number d of hidden units can be varied in our analysis. For notational simplicity, we incorporate the bias b in the weight $\boldsymbol{w}$ as $\boldsymbol{w} = (b, w^1, \ldots, w^n)$, and accordingly, we enlarge $\boldsymbol{x}$ as $\boldsymbol{x} = (1, x_1, \ldots, x_n)$. By using these conventions, we obtain the abridged presentation of the three-layer perceptron as

$$\boldsymbol{f}_{(d)}(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{i=1}^{d} \boldsymbol{v}_i \, \varphi\left(\boldsymbol{w}_i \cdot \boldsymbol{x}\right). \tag{2}$$

Figure 1 is a schematic diagram of the three-layer perceptron.

In this paper, we treat the *supervised learning*, which aims at finding a parameter $\boldsymbol{\theta}$ so that $\boldsymbol{f}_{(d)}(\boldsymbol{x};\boldsymbol{\theta})$ approximates a given target function $T(\boldsymbol{x})$. The *(averaged) gradient descent method* is a standard method to find such $\boldsymbol{\theta}$ numerically. Suppose that a loss function $\ell(\boldsymbol{x}, \boldsymbol{y})$ is non-negative and is equal to zero if and only if $\boldsymbol{y} = T(\boldsymbol{x})$ (*e.g.* the squared error $||\boldsymbol{y} - T(\boldsymbol{x})||^2$). In the gradient descent method, we aim at minimising the averaged loss function

$$L_{(d)}(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{x}}\left[\ell(\boldsymbol{x}, \boldsymbol{f}_{(d)}(\boldsymbol{x};\boldsymbol{\theta}))\right] \tag{3}$$

by changing the parameter $\boldsymbol{\theta}$ according to the gradient system

$$\frac{d\boldsymbol{\theta}}{dt} = -\frac{\partial L_{(d)}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}). \tag{4}$$

The parameter $\boldsymbol{\theta}$ descends along the gradient of $L_{(d)}$ to reach a local minimiser. Here, we assume that the input vector $\boldsymbol{x}$ is a random variable drawn according to an unknown probability distribution, and $\mathbb{E}_{\boldsymbol{x}}$ denotes the expectation with respect to $\boldsymbol{x}$.

1.2 Singular region and Milnor-like attractor

The parameter space of a hierarchical neural network usually contains a subset whose points correspond to the same input-output relation. Such a subset is referred to as a *singular region*. For example, let us consider an (n-2-m)-perceptron. Then, for $\boldsymbol{w} \in \mathbb{R}^{n+1}, \boldsymbol{v} \in \mathbb{R}^m$, the subset

$$R(\boldsymbol{w}, \boldsymbol{v}) := \{\, \boldsymbol{\theta} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{v}_1, \boldsymbol{v}_2) \mid \boldsymbol{w}_1 = \boldsymbol{w}_2 = \boldsymbol{w}, \boldsymbol{v}_1 + \boldsymbol{v}_2 = \boldsymbol{v} \,\}$$

is a singular region. In fact, on the subset $R(\boldsymbol{w}, \boldsymbol{v})$, an ($n$-2-$m$)-perceptron $\boldsymbol{f}_{(2)}(\boldsymbol{x}; \boldsymbol{\theta})$ is reduced to a (n-1-m)-perceptron as

$$\boldsymbol{f}_{(1)}(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{v}) = \boldsymbol{v}\, \varphi\, (\boldsymbol{w} \cdot \boldsymbol{x})\,.$$

On such a singular region, some properties of $L_{(1)}$ are inherited by $L_{(2)}$. The following theorem holds, for example.

Theorem 1.1 ([2], Theorem 1)**.** *Let $\boldsymbol{\theta}^* = (\boldsymbol{w}^*, \boldsymbol{v}^*)$ be a critical point of $L_{(1)}$. Then, the parameter $\boldsymbol{\theta} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{v}_1, \boldsymbol{v}_2) = (\boldsymbol{w}^*, \boldsymbol{w}^*, \lambda \boldsymbol{v}^*, (1-\lambda)\boldsymbol{v}^*)$ is a critical point of $L_{(2)}$ for any $\lambda \in \mathbb{R}$.*

When $m = 1$, in particular, every point $\boldsymbol{\theta} \in R(\boldsymbol{w}^*, v^*)$ is a critical point of $L_{(2)}$. Further, in this case, the second order property of $L_{(1)}$ is also inherited by $L_{(2)}$ to some extent, and the singular region $R(\boldsymbol{w}^*, v^*)$ may have an interesting structure which causes serious stagnation of learning.

Theorem 1.2 ([2], Theorem 3)**.** *Let $m = 1$ and $\boldsymbol{\theta}^* = (\boldsymbol{w}^*, v^*)$ be a strict local minimiser of $L_{(1)}$ with $v^* \neq 0$. Define an $(n+1) \times (n+1)$ matrix*

$$H := \mathbb{E}_{\boldsymbol{x}} \left[\frac{\partial \ell(\boldsymbol{x}, f_{(1)}(\boldsymbol{x}; \boldsymbol{\theta}^*))}{\partial y}\, v^* \varphi''(\boldsymbol{w}^* \cdot \boldsymbol{x})\, \boldsymbol{x}\, \boldsymbol{x}^T \right], \tag{5}$$

and for $\lambda \in \mathbb{R}$

$$\boldsymbol{\theta}_\lambda := (\boldsymbol{w}^*, \boldsymbol{w}^*, \lambda v^*, (1-\lambda) v^*).$$

If the matrix H is positive (resp. negative) definite, then $\boldsymbol{\theta} = \boldsymbol{\theta}_\lambda$ is a local minimiser (resp. saddle point) of $L_{(2)}$ for any $\lambda \in (0,1)$, and is a saddle point (resp. local minimiser) for any $\lambda \in \mathbb{R} \setminus [0,1]$. On the other hand, if the matrix H is indefinite, then the point $\boldsymbol{\theta}_\lambda$ is a saddle point of $L_{(2)}$ for all $\lambda \in \mathbb{R} \setminus \{0,1\}$.

This theorem implies that the one-dimensional region $R(\boldsymbol{w}^*, v^*) = \{\boldsymbol{\theta}_\lambda \mid \lambda \in \mathbb{R}\}$ may have both attractive parts and repulsive parts in the gradient descent method. Such a region is referred to as a *Milnor-like attractor* [4]. The parameter $\boldsymbol{\theta}$ near the attractive part flows into the Milnor-like attractor; however, since there are some stochastic effects in practical learning, it fluctuates around the Milnor-like attractor. When it reaches the repulsive part by such fluctuation, the parameter escapes from the Milnor-like attractor, and the loss starts to decrease again. Figure 2 is a schematic diagram of a Milnor-like attractor.

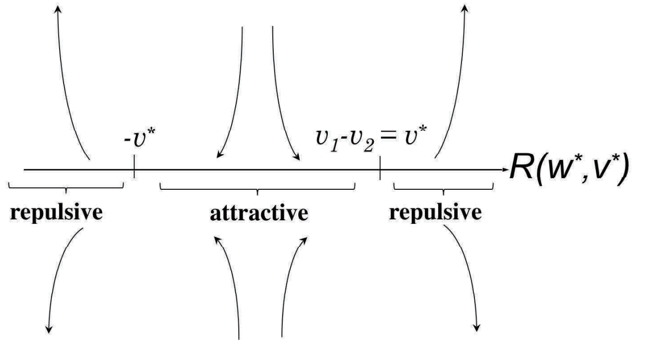


Figure 2: A schematic diagram of a Milnor-like attractor $R(\boldsymbol{w}^*, v^*)$. A parameter fluctuates around the attractive part of a Milnor-like attractor for a long time by some stochastic effects, until it reaches the repulsive part.

When $m \geq 2$, there also exists a one-dimensional region consisting of critical points due to Theorem 1.1; however, the region becomes simply repulsive, and does not have an attractive part as the following theorem asserts.

Theorem 1.3 ([3], Theorem 2.3)**.** *Let* $\boldsymbol{\theta}^* = (\boldsymbol{w}^*, \boldsymbol{v}^*)$ *be a local minimiser of* $L_{(1)}$*. If the* $m \times (n+1)$ *matrix*

$$\mathbb{E}_{\boldsymbol{x}}\left[\frac{\partial \ell(\boldsymbol{x}, \boldsymbol{f}_{(1)}(\boldsymbol{x}; \boldsymbol{\theta}^*))}{\partial \boldsymbol{y}} \varphi'(\boldsymbol{w}^* \cdot \boldsymbol{x})\, \boldsymbol{x}^T\right]$$

is non-zero, then $\boldsymbol{\theta}_\lambda = (\boldsymbol{w}^*, \boldsymbol{w}^*, \lambda \boldsymbol{v}^*, (1-\lambda)\boldsymbol{v}^*)$ *is a saddle point of* $L_{(2)}$ *for any* $\lambda \in \mathbb{R}$*, where we regard the derivative* $\partial \ell / \partial \boldsymbol{y}$ *as a column vector.*

In their article [1], Amari *et al.* stated a prototype of Theorem 1.3.

2 Centre Manifold of Milnor-like Attractor

In their analysis of an (n-2-1)-perceptron, Wei *et al.* [4] introduced a coordinate transformation of the parameter space

$$\begin{cases} \boldsymbol{w} = \dfrac{v_1 \boldsymbol{w}_1 + v_2 \boldsymbol{w}_2}{v_1 + v_2} \\ v = v_1 + v_2 \\ \boldsymbol{u} = \boldsymbol{w}_1 - \boldsymbol{w}_2 \\ z = \dfrac{v_1 - v_2}{v_1 + v_2} \end{cases} \tag{6}$$

and claimed, based on evidences found in numerical simulations, that the parameters $(\boldsymbol{w}, v)$ quickly converge to $(\boldsymbol{w}^*, v^*)$ when the initial point is taken near a Milnor-like attractor. Amari *et al.* [1] mentioned that the dynamics in this coordinate system should be analysed by using the centre manifold theory, and they analysed only the reduced dynamical system for the sub-parameters $(\boldsymbol{u}, z)$, setting the remaining parameters $(\boldsymbol{w}, v)$ to be $(\boldsymbol{w}^*, v^*)$.

While the coordinate system (6), in fact, does not admit any centre manifold structure, it is the case that there exists a coordinate system that admits a centre manifold structure. Such a coordinate system $\boldsymbol{\xi} = (\boldsymbol{w}, v, \boldsymbol{u}, z)$ is, for example, given as follows.

$$\begin{cases} \boldsymbol{w} = \dfrac{v_1 (\boldsymbol{w}_1 - \boldsymbol{w}^*) + v_2 (\boldsymbol{w}_2 - \boldsymbol{w}^*)}{v^*} + \boldsymbol{w}^* \\ v = v_1 + v_2 \\ \boldsymbol{u} = \dfrac{v_2 (\boldsymbol{w}_1 - \boldsymbol{w}^*) - v_1 (\boldsymbol{w}_2 - \boldsymbol{w}^*)}{v^*} \\ z = v_1 - v_2 \end{cases} . \tag{7}$$

This formula defines a coordinate system on the region $\{v_1^2 + v_2^2 \neq 0\}$. Now the following theorem holds.

Theorem 2.1 ([3], Theorem 3.5)**.** *In the coordinate system $\boldsymbol{\xi} = (\boldsymbol{w}, v, \boldsymbol{u}, z)$, the dynamical system (4) admits a centre manifold structure around the critical points $\boldsymbol{\theta} = \boldsymbol{\theta}_0, \boldsymbol{\theta}_1$ in which $(\boldsymbol{w}, v)$ converge exponentially fast.*

Let us remark that the two points $\boldsymbol{\theta} = \boldsymbol{\theta}_0, \boldsymbol{\theta}_1$ are the boundaries of repulsive and attractive parts of the Milnor-like attractor $\{ \boldsymbol{\theta}_\lambda \mid \lambda \in \mathbb{R} \}$. Thus, when passing nearby these points, a parameter evolving around the Milnor-like attractor changes the mode of dynamics. Due to this theorem, we can perform a detailed analysis by using the centre manifold reduction.

Due to the standard method from the centre manifold theory, we obtain the reduced dynamical system around $\boldsymbol{\theta} = \boldsymbol{\theta}_1$ as

$$\begin{aligned} \dot{\boldsymbol{u}} &= \frac{1}{2v^*}(z - v^*)H\boldsymbol{u} + O(\|\boldsymbol{u}, z - v^*\|^3), \\ \dot{z} &= \frac{1}{2v^*}\boldsymbol{u}^T H\boldsymbol{u} + O(\|\boldsymbol{u}, z - v^*\|^3). \end{aligned} \tag{8}$$

Note that the point $\boldsymbol{\theta} = \boldsymbol{\theta}_1$ is denoted as $\boldsymbol{\xi} = \boldsymbol{\xi}_1 = (\boldsymbol{w}^*, v^*, \mathbf{0}, v^*)$ under the coordinate system (7). Neglecting the higher order terms, we can integrate this equation to obtain

$$\|\boldsymbol{u}\|^2 = (z - v^*)^2 + C, \tag{9}$$

where C is an integral constant.

Around the point $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, or equivalently $\boldsymbol{\xi} = \boldsymbol{\xi}_0 = (\boldsymbol{w}^*, v^*, \mathbf{0}, -v^*)$, the similar reduced dynamical system:

$$\begin{aligned} \dot{\boldsymbol{u}} &= -\frac{1}{2v^*}(z + v^*)H\boldsymbol{u} + O(\|\boldsymbol{u}, z + v^*\|^3), \\ \dot{z} &= -\frac{1}{2v^*}\boldsymbol{u}^T H\boldsymbol{u} + O(\|\boldsymbol{u}, z + v^*\|^3), \end{aligned}$$

is obtained.

3 Numerical simulations

We shall verify the fact that the dynamics of $(\boldsymbol{w}, v)$ are fast and those of $(\boldsymbol{u}, z)$ are slow under the coordinate system (7) by numerical simulations.

We set the input dimension to be $n = 1$, and choose the teacher function $T : \mathbb{R} \to \mathbb{R}$ defined by

$$T(x) := 2\tanh(x) - \tanh(4x).$$

We set the activation function φ as tanh. Thus, the target function T can be represented by the (1-2-1)-perceptron with no bias terms:

$$f_{(2)}(x; \boldsymbol{\theta}) = v_1\varphi(w_1 x) + v_2\varphi(w_2 x),$$

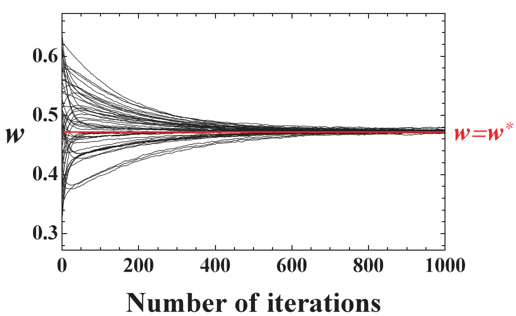
and the true parameter is $(w_1, w_2, v_1, v_2) = (1, 4, 2, -1)$. We also discard the bias terms of the student (1-1-1)-perceptron. This makes the matrix H defined by (5) scalar valued, and it becomes positive or negative definite trivially.

We set the probability distribution of the input x to be the Gaussian distribution $N(0, 2^2)$. Taking a large size of dataset $\{x_s\}_{s=1}^{S}$ according to $N(0, 2^2)$ for each iteration, we compute the arithmetic mean of the instantaneous loss $\ell(x, f_{(2)}(x; \boldsymbol{\theta}))$ over $\{x_s\}_{s=1}^{S}$, which approximates the averaged loss function (3). Thus, the transition formula of the parameter $\boldsymbol{\theta}$ is written as

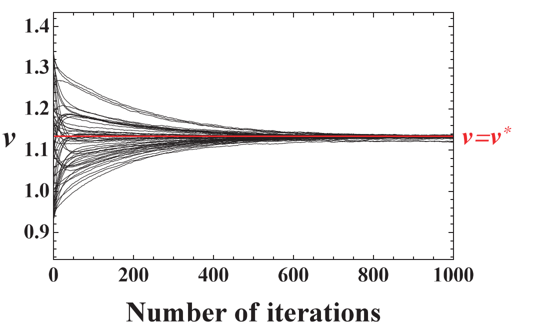
$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \varepsilon \frac{1}{S} \sum_{s=1}^{S} \left. \frac{\partial \ell(x_s^{(t)}, f_{(2)}(x_s^{(t)}; \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} . \tag{10}$$

Here, $\varepsilon > 0$ is a small number for the Euler method. In this simulation, we set $S = 500$, $\varepsilon = 0.05$, and the loss function ℓ to be the squared error.

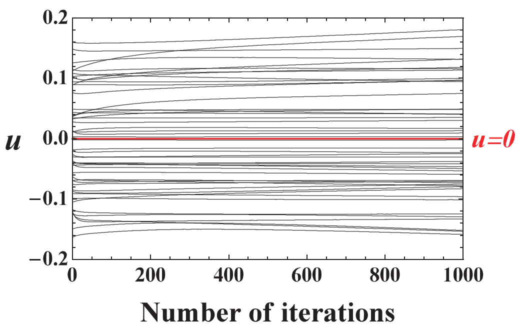
In this setting, we obtained a local minimiser $\boldsymbol{\theta}^* = (w^*, v^*) \approx (0.472, 1.134)$ of $L_{(1)}$. The value of H is approximately 0.050. Since $H > 0$, the attractive region is $\{\boldsymbol{\theta}_\lambda \mid \lambda \in (0, 1)\}$, due to Theorem 1.2.



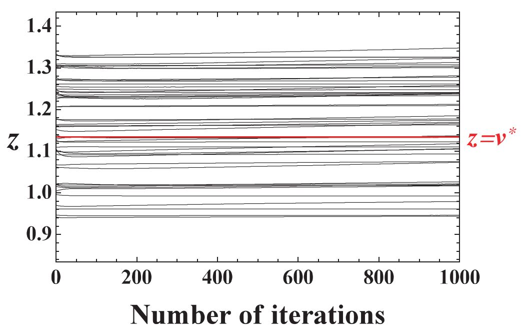
(a) Time evolutions of w.



(b) Time evolutions of v.



(c) Time evolutions of u.



(d) Time evolutions of z.

Figure 3: Time evolutions of each parameter in the first 1,000 iterations. Each trajectory of (w, v) quickly converges to $(w^*, v^*) \approx (0.472, 1.134)$, while trajectories of u and z evolve very slowly.

Figures 3(a-d) display time evolutions of each parameter in the first 1,000 iterations from 50 different initial points. We chose an initial parameter $\boldsymbol{\theta}^{(0)} = (w_1^{(0)}, w_2^{(0)}, v_1^{(0)}, v_2^{(0)})$ randomly by

$$w_1^{(0)} = w^* + \zeta_1, \quad w_2^{(0)} = w^* + \zeta_2,$$
$$v_1^{(0)} = v^* + \frac{1}{2}(\zeta_3 + \zeta_4), \quad v_2^{(0)} = \frac{1}{2}(\zeta_3 - \zeta_4),$$

so that $v = v^* + \zeta_3$, and $z = v^* + \zeta_4$, where $\zeta_1, \zeta_2 \sim U(-0.2, 0.2)$, and $\zeta_3, \zeta_4 \sim U(-0.2, 0.2)$. Here, $U(a, b)$ denotes the uniform distribution on the interval $[a, b] \subset \mathbb{R}$. We can see that the parameters w and v converge to their equilibriums exponentially fast ((a) and (b)), while u and z evolve slowly ((c) and (d)).

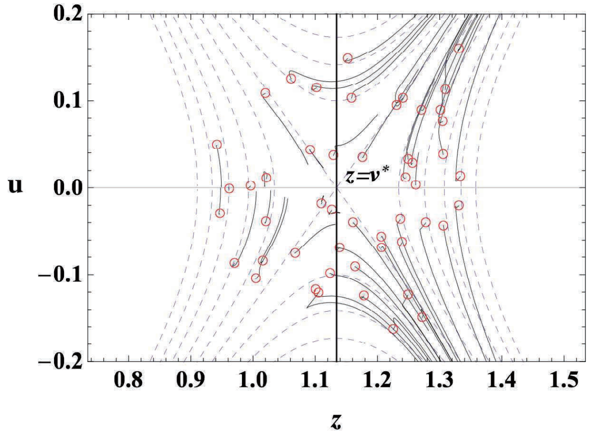


Figure 4: Trajectories on the (z, u)-plane obtained by learning for 20,000 iterations (solid black curves) and analytical trajectories (9) (dashed blue curves) near $\boldsymbol{\theta} = \boldsymbol{\theta}_1 = (w^*, w^*, v^*, 0)$. Red circles represent initial points.

Figure 4 shows evolutions on the (z, u)-plane. The red circles in the figure represent initial points. When $(w, v) = (w^*, v^*)$, the z-axis is a Milnor-like attractor, and the region $|z| < v^*$ is the attractive part of it. The intersection point of the line $z = v^*$ and the z-axis corresponds to the point $\boldsymbol{\theta} = \boldsymbol{\theta}_1$. The analytical trajectories (9) are plotted as dashed blue curves. Numerical evolutions of the parameter follow the analytical trajectories considerably well around $\boldsymbol{\theta}_1$.

References

[1] S.-I. Amari, T. Ozeki, R. Karakida, Y. Yoshida and M. Okada, "Dynamics of learning in MLP: Natural gradient and singularity revisited," Neural Computation **30**(1), 1-33 (2018).

[2] K. Fukumizu and S.-I. Amari, "Local minima and plateaus in hierarchical structures of multilayer perceptrons," Neural Networks **13**(3), 317–327 (2000).

[3] D. Tsutsui, "Centre manifold analysis of plateau phenomena caused by degeneration of three-layer perceptron," Neural Computation **32**(4), 683–710 (2020).

[4] H. Wei, J. Zhang, F. Cousseau, T. Ozeki, and S.-I. Amari, "Dynamics of learning near singularities in layered networks," Neural Computation **20**(3), 813–843 (2008).