ARTICLE **Communicated by Shun-ichi Amari**

Center Manifold Analysis of Plateau Phenomena Caused by Degeneration of Three-Layer Perceptron

Daiji Tsutsui
d-tsutsui@cr.math.sci.osaka-u.ac.jp
Osaka University, Toyonaka-shi, Osaka 560-0043, Japan

A hierarchical neural network usually has many singular regions in the parameter space due to the degeneration of hidden units. Here, we focus on a three-layer perceptron, which has one-dimensional singular regions comprising both attractive and repulsive parts. Such a singular region is often called a Milnor-like attractor. It is empirically known that in the vicinity of a Milnor-like attractor, several parameters converge much faster than the rest and that the dynamics can be reduced to smaller-dimensional ones. Here we give a rigorous proof for this phenomenon based on a center manifold theory. As an application, we analyze the reduced dynamics near the Milnor-like attractor and study the stochastic effects of the online learning.

1 Introduction

A three-layer perceptron is one of the simplest hierarchical learning machines. Mathematically, it is defined by the function

$$f_{(d)}(x;\boldsymbol{\theta}) = \sum_{i=1}^{d} \boldsymbol{v}_i \varphi\left(\boldsymbol{w}_i \cdot \boldsymbol{x} + b_i\right), \quad \boldsymbol{x} \in \mathbb{R}^n, \tag{1.1}$$

$$\boldsymbol{\theta} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_d, b_1, \ldots, b_d, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_d),$$

where $\boldsymbol{\theta}$ is a system parameter with $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_d \in \mathbb{R}^n$ being the weight vectors for the second layer, $b_1, \ldots, b_d \in \mathbb{R}$ the bias terms for the second, $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d \in \mathbb{R}^m$ the weight vectors for the third, and φ an activation function. Throughout this article, we assume that the activation function is twice differentiable. Figure 1 is a schematic diagram of the three-layer perceptron. We shall call the function 1.1 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

Neural Computation **32, 683–710** (2020)
https://doi.org/10.1162/neco_a_01268

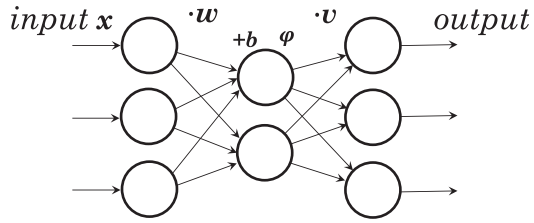


Figure 1: A schematic diagram of a three-layer perceptron presented in equation 1.1.

$$f_{(d)}(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{i=1}^{d} \boldsymbol{v}_i\, \varphi\,(\boldsymbol{w}_i \cdot \boldsymbol{x})\,. \tag{1.2}$$

In this article, we treat the supervised learning. The goal of the supervised learning is to find an optimal parameter $\boldsymbol{\theta}$ so that $f_{(d)}(\boldsymbol{x};\boldsymbol{\theta})$ approximates a given target function $T(\boldsymbol{x})$. Such a problem is based on the universal approximation property of the three-layer perceptron. For a suitable activation function φ (e.g., sigmoidal or ReLU), the model, equation 1.2, can approximate quite a wide range of functions as the number d of hidden units tends to infinity (Cybenko, 1989; Funahashi, 1989; Sonoda & Murata, 2017).

The (averaged) gradient descent method is a standard method for finding an optimal value of $\boldsymbol{\theta}$ numerically. Suppose that a loss function $\ell(\boldsymbol{x},\boldsymbol{y})$ is nonnegative 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 minimizing

$$L_{(d)}(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{x}}\left[\ell(\boldsymbol{x},\, f_{(d)}(\boldsymbol{x};\boldsymbol{\theta}))\right] \tag{1.3}$$

by changing the parameter $\boldsymbol{\theta}$ iteratively as

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \varepsilon \frac{\partial L_{(d)}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_t), \tag{1.4}$$

where $\varepsilon > 0$ is a learning constant. 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 $\boldsymbol{x}$. In order for the differential $\partial L_{(d)}/\partial\boldsymbol{\theta}$ to make sense, we also assume that $\ell(\boldsymbol{x},\boldsymbol{y})$ is differentiable with respect to $\boldsymbol{y}$ and that we can interchange the order of the differentiation $\partial/\partial\boldsymbol{\theta}$ and the expectation $\mathbb{E}_{\boldsymbol{x}}$. We study the dynamical system, which represents the averaged gradient descent method with infinitesimal learning constant:

$$\frac{d\boldsymbol{\theta}}{dt} = -\frac{\partial L_{(d)}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}). \tag{1.5}$$

The parameter $\boldsymbol{\theta}$ descends along the gradient of $L_{(d)}$ into a local minimum. In practice, the expectation in equation 1.3 is replaced with the arithmetic mean over a given data set, or, roughly, with a single realization of the random variable $\ell(\boldsymbol{x}, \boldsymbol{f}_{(d)}(\boldsymbol{x}; \boldsymbol{\theta}))$ for each learning iteration. Such a learning method involving some stochastic effects is called a *stochastic gradient descent method*.

Fukumizu and Amari (2000) studied singular regions arising from degeneration of hidden units of a three-layer perceptron. Here, the degeneration of hidden units means that several weight parameters $\boldsymbol{w}_i$ take the same value and the effective number of hidden units becomes fewer than d. When $m = 1$, they found a novel type of singular region, often called a *Milnor-like attractor*. This region has both an attractive part consisting of local minima of $L_{(d)}$ and a repulsive part consisting of saddle points. In practical learning, there may be some stochastic effects. Therefore, once the parameter $\boldsymbol{\theta}$ is trapped in the attractive part of this region, it fluctuates in the region by stochastic effects for a long time, until it reaches the repulsive part. This may cause serious stagnation of learning, called *plateau phenomena*. Later, Amari, Ozeki, Karakida, Yoshida, and Okada (2018) pointed out a notable fact that a Milnor-like attractor may not cause serious stagnation of learning when $m \geq 2$, which is also treated in this article.

More quantitative analyses for $m = 1$ have also been carried out by Coussear, Ozeki, and Amari (2008), Wei, Zhang, Cousseau, Ozeki, and Amari (2008), and Amari et al. (2018) in the simplest case $d = 2$. In particular, Wei et al. (2008) introduced a new coordinate system in the parameter space by

$$\begin{cases} \boldsymbol{w} = \dfrac{v_1 \boldsymbol{w}_1 + v_2 \boldsymbol{w}_2}{v_1 + v_2} \\ v = v_1 + v_2 \\ \boldsymbol{u} = \boldsymbol{w}_1 - \boldsymbol{w}_2 \\ z = \dfrac{v_1 - v_2}{v_1 + v_2} \end{cases}, \tag{1.6}$$

and claimed, based on evidence observed in numerical simulations, that when the initial point is taken near a Milnor-like attractor, the parameters $(\boldsymbol{w}, v)$ quickly converge to equilibrium values $(\boldsymbol{w}^*, v^*)$. They hypothesized that this would always be the case and analyzed only the reduced dynamical system for the subparameters $(\boldsymbol{u}, z)$, setting the remaining parameters $(\boldsymbol{w}, v)$ to be $(\boldsymbol{w}^*, v^*)$. However, to the best of our knowledge, no mathematical justification for this hypothesis has been established.

The objective of this article is to provide a solid ground on Amari et al.'s (2018) point of view. We introduce a new coordinate system that admits a center manifold structure around a special point on the Milnor-like attractor. By using the coordinate system, we can analyze the Milnor-like

attractor more rigorously and integrate the reduced dynamical system explicitly to obtain analytical trajectories. The obtained trajectories are comparable to the preceding work. It is confirmed by several settings of numerical simulations that trajectories in actual learning agree with the analytical ones.

In addition to the averaged gradient descent method, we also address online learning, a stochastic gradient descent method. Around a Milnor-like attractor, the behavior of sample paths by the online learning seems qualitatively different from that of trajectories by the averaged gradient descent. To investigate why they are different, we divide the dynamics of parameters into fast and slow ones, as is the case in the averaged gradient descent. In this case, we observed in numerical simulations that the fast parameters fluctuate intensively around the center manifold for the averaged system. We show that such a deviation of the fast parameter from the center manifold can influence a trend of the slow parameter.

This article is organized as follows. In section 2, we give a quick review of Amari et al.'s (2018) work. In section 3, after a brief account of the center manifold theory, we introduce a new coordinate system in the parameter space and prove that it admits the center manifold structure. In section 4, we carry out numerical simulations and observe the center manifold structure around a Milnor-like attractor. In section 5, we consider the online learning from the viewpoint of the center manifold theory. Section 6 offers concluding remarks.

2 Singular Region and Milnor-Like Attractor

In this section, we give a quick review of the Milnor-like attractor that Fukumizu and Amari (2000) found, which appears when the number m of output units is equal to 1. We also mention an interesting insight by Amari et al. (2018) for the case $m \geq 2$.

The parameter space of a perceptron is sometimes called a perceptron manifold. However, in many cases, it is not really a manifold since it usually contains a subset whose points correspond to the same input-output relation. Such a subset is usually referred to as a singular region. In general, there are many singular regions due to the degeneration of hidden units. For example, let us consider an (n-2-m)-perceptron. Then the subset

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

of the parameter space forms a typical 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 the following (n-1-m)-perceptron:

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

On such a singular region, some properties of $L_{(1)}$ are inherited by $L_{(2)}$. The following lemma implies that a criticality is a hereditary property.

Lemma 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 $\lambda \in \mathbb{R}$.*

Proof.

$$\frac{\partial L_{(2)}}{\partial \boldsymbol{w}_i}(\boldsymbol{\theta}) = \mathbb{E}\left[\left(\frac{\partial \ell(\boldsymbol{x}, \boldsymbol{f}_{(1)}(\boldsymbol{x}; \boldsymbol{\theta}^*))}{\partial \boldsymbol{y}} \cdot \boldsymbol{v}_i\right) \varphi'(\boldsymbol{w}^* \cdot \boldsymbol{x}) \boldsymbol{x}^T\right] = \lambda_i \frac{\partial L_{(1)}}{\partial \boldsymbol{w}}(\boldsymbol{\theta}^*),$$

$$\frac{\partial L_{(2)}}{\partial \boldsymbol{v}_i}(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{\partial \ell(\boldsymbol{x}, \boldsymbol{f}_{(1)}(\boldsymbol{x}; \boldsymbol{\theta}^*))}{\partial \boldsymbol{y}} \varphi(\boldsymbol{w}^* \cdot \boldsymbol{x})\right] = \frac{\partial L_{(1)}}{\partial \boldsymbol{v}}(\boldsymbol{\theta}^*), \qquad i = 1, 2,$$

where $\lambda_1 := \lambda$ and $\lambda_2 := 1 - \lambda$. Since $\boldsymbol{\theta}^*$ is a critical point of $L_{(1)}$, these are all zero. □

When $m = 1$, in particular, every point $\boldsymbol{\theta} \in R(\boldsymbol{w}^*, v^*)$ is a critical point of $L_{(2)}$, since the parameter v, as well as the output $f^{(d)}(\boldsymbol{x}; \boldsymbol{\theta})$, is a scalar quantity. 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.

Proposition 1 *(Fukumizu & Amari, 2000). Let $m = 1$ and $\boldsymbol{\theta}^* = (\boldsymbol{w}^*, v^*)$ be a strict local minimizer of $L_{(1)}$ with $v^* \neq 0$. Define an $(n+1) \times (n+1)$ matrix H by*

$$H := \mathbb{E}_{\boldsymbol{x}}\left[\frac{\partial \ell(\boldsymbol{x}, f_{(1)}(\boldsymbol{x}; \boldsymbol{\theta}^*))}{\partial y} v^* \varphi''(\boldsymbol{w}^* \cdot \boldsymbol{x}) \boldsymbol{x} \boldsymbol{x}^T\right], \tag{2.1}$$

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

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

If the matrix H is positive (resp. negative) definite, then the point $\boldsymbol{\theta} = \boldsymbol{\theta}_\lambda$ is a local minimizer (resp. saddle point) of $L_{(2)}$ for any $\lambda \in (0, 1)$ and is a saddle point (resp. local minimizer) 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 proposition 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* (Wei et al., 2008). The parameter $\boldsymbol{\theta}$ near the attractive part flows into the Milnor-like attractor and fluctuates in the region for a long time, until it reaches the repulsive part.

The original theorem (Fukumizu & Amari, 2000) is for an (n-d-1)-perceptron that contains (n-(d-1)-1)-perceptron as a subnetwork and that

the phenomenon itself is universal with respect to the number d of hidden units. The proposition above for (n-2-1)-perceptron is a minimal version.

We also remark that the point $\boldsymbol{\theta}_\lambda$ cannot be a strict local minimizer since $L_{(2)}$ takes the same value on the singular region $\{\boldsymbol{\theta}_\lambda \mid \lambda \in \mathbb{R}\}$ and is flat along its direction. The proof of proposition 1 is given mainly by a discussion of the Hessian matrix of $L_{(2)}$; however, we need to treat higher-order derivatives of $L_{(2)}$ than the second order, since the Hessian matrix degenerates on the singular region (see appendix A).

Let us suppose a situation where a three-layer perceptron has some redundant hidden units to represent the target function $T(\boldsymbol{x})$. Mathematically, we suppose that a true parameter $\boldsymbol{\theta}_{true}$ exists ($T(\boldsymbol{x}) = f_{(2)}(\boldsymbol{x}; \boldsymbol{\theta}_{true})$) and that it lies in the singular region $R(\boldsymbol{w}^*, v^*)$. In this case, the function $L_{(2)}$ takes the same value $L_{(1)}(\boldsymbol{w}^*, v^*) = 0$ on $R(\boldsymbol{w}^*, v^*)$. Therefore, every point of $R(\boldsymbol{w}^*, v^*)$ becomes a global minimizer of $L_{(2)}$, and a Milnor-like attractor does not appear. In fact, one can check that the assumption of proposition 1 fails as follows. For each $\boldsymbol{x} \in \mathbb{R}^n$, we obtain

$$\partial \ell(\boldsymbol{x}, f_{(1)}(\boldsymbol{x}; \boldsymbol{w}^*, v^*))/\partial y = 0,$$

since a loss function $\ell(\boldsymbol{x}, \boldsymbol{y})$ takes its minimum 0 at $y = T(\boldsymbol{x}) = f_{(1)}(\boldsymbol{x}; \boldsymbol{w}^*, v^*)$. This implies that the matrix H becomes the zero matrix. Thus, H is neither positive nor negative definite.

We next treat the case when $m \geq 2$. There also exists a one-dimensional region consisting of critical points due to lemma 1. However, in this case, the region becomes simply repulsive and does not have an attractive part, as the following theorem asserts.

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

$$\mathbb{E}_x \left[\frac{\partial \ell(\boldsymbol{x}, \boldsymbol{f}_{(1)}(\boldsymbol{x}; \boldsymbol{\theta}^*))}{\partial \boldsymbol{y}} \varphi'(\boldsymbol{w}^* \cdot \boldsymbol{x}) \boldsymbol{x}^T \right] \tag{2.2}$$

is nonzero, 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.

Amari et al. (2018) stated a prototype of theorem 1, although they did not give a full proof. In fact, we found that some additional assumption was necessary to prove their assertion. In theorem 1, we have added a mild assumption that the matrix, equation 2.2, is nonzero. Note that since $\boldsymbol{\theta}^* = (\boldsymbol{w}^*, \boldsymbol{v}^*)$ is a local minimizer of $L_{(1)}$, it holds that

$$\mathbf{0} = \frac{\partial L_{(1)}}{\partial \boldsymbol{w}}(\boldsymbol{\theta}^*) = (\boldsymbol{v}^*)^T \, \mathbb{E}_{\boldsymbol{x}} \left[\frac{\partial \ell(\boldsymbol{x}, \boldsymbol{f}_{(1)}(\boldsymbol{x}; \boldsymbol{\theta}^*))}{\partial \boldsymbol{y}} \varphi'(\boldsymbol{w}^* \cdot \boldsymbol{x}) \boldsymbol{x}^T \right].$$

Thus, the matrix 2.2 has a kernel whose dimension is greater than or equal to one. Hence, the assumption automatically fails when $m = 1$. This is an underlying mechanism for proposition 1.

3 Center Manifold of Milnor-Like Attractor

In their analysis of an (n-2-1)-perceptron, Wei et al. (2008) introduced a coordinate transformation,

$$\begin{cases} \boldsymbol{w} = \dfrac{v_1 \boldsymbol{w}_1 + v_2 \boldsymbol{w}_2}{v_1 + v_2} \\ v = v_1 + v_2 \\ \boldsymbol{u} = \boldsymbol{w}_1 - \boldsymbol{w}_2 \\ z = \dfrac{v_1 - v_2}{v_1 + v_2} \end{cases}, \tag{3.1}$$

and claimed 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. (2018) mentioned that the dynamics in this coordinate system should be analyzed by using the center manifold theory, and they analyzed only the reduced dynamical system for the subparameters $(\boldsymbol{u}, z)$, setting the remaining parameters $(\boldsymbol{w}, v)$ to be $(\boldsymbol{w}^*, v^*)$. Strictly speaking, however, their coordinate system does not admit any center manifold structure, and their claim is at the stage of hypothesis.

In this section, we give a rigorous justification for their hypothesis. We first give a quick review of the center manifold theory and then introduce a new coordinate system under which the center manifold structures arise near certain points on the Milnor-like attractor.

3.1 Brief Review of Center Manifold. Suppose that we are given a dynamical system,

$$\begin{cases} \dot{\boldsymbol{x}}(t) = A\boldsymbol{x}(t) + \boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{y}(t)) \\ \dot{\boldsymbol{y}}(t) = B\boldsymbol{y}(t) + \boldsymbol{g}(\boldsymbol{x}(t), \boldsymbol{y}(t)) \end{cases}, \tag{3.2}$$

for the parameters $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, where A and B are constant matrices and $\boldsymbol{f}$ and $\boldsymbol{g}$ are C^2 functions such that they, along with their first derivatives, vanish at the origin. We assume that all the eigenvalues of A have zero real parts, while all the eigenvalues of B have negative real parts. This assumption means that parameter $\boldsymbol{y}$ converges to the origin exponentially quickly, and the parameter $\boldsymbol{x}$ is driven only by the higher-order terms of $\boldsymbol{f}$

and evolves very slowly compared with $\boldsymbol{y}$. Since $\boldsymbol{f}$ and $\boldsymbol{g}$ are of the second order with respect to $\boldsymbol{x}$ and $\boldsymbol{y}$, the assumptions for equation 3.2 imply that the coefficient matrix of the linearization of the system has the form

$$\begin{pmatrix} A & O \\ O & B \end{pmatrix}.$$

Definition 1. *A set $S \subset \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ is said to be a local invariant manifold of equation 3.2 if for $(\boldsymbol{x}_0, \boldsymbol{y}_0) \in S$, the solution $(\boldsymbol{x}(t), \boldsymbol{y}(t))$ of equation 3.2 with $(\boldsymbol{x}(0), \boldsymbol{y}(0)) = (\boldsymbol{x}_0, \boldsymbol{y}_0)$ is in S for $|t| < T$ with some $T > 0$.*

Definition 2. *A local invariant manifold represented in the form of $\boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x})$ is called a local center manifold (or simply a center manifold) if $\boldsymbol{h}$ is differentiable and satisfies $\boldsymbol{h}(\boldsymbol{0}) = \boldsymbol{0}$ and $\dfrac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}}(\boldsymbol{0}) = O$.*

The following center manifold theorems give us a method of simplifying a dynamical system around an equilibrium point.

Proposition 2 *(Center manifold theorem 1: Carr, 1981). Equation 3.2 has a center manifold $\boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x})$ for $||\boldsymbol{x}|| < \delta$, for some $\delta > 0$ and C^2 function $\boldsymbol{h}$.*

Proposition 3 *(Center manifold theorem 2: Carr, 1981). Suppose that the origin $\boldsymbol{u} = \boldsymbol{0}$ is a stable equilibrium point of the reduced dynamical system*

$$\dot{\boldsymbol{u}}(t) = A\boldsymbol{u}(t) + \boldsymbol{f}(\boldsymbol{u}(t), \boldsymbol{h}(\boldsymbol{u}(t))). \tag{3.3}$$

Let $(\boldsymbol{x}(t), \boldsymbol{y}(t))$ be a solution of equation 3.2 with the initial value $(\boldsymbol{x}_0, \boldsymbol{y}_0)$. Then if $||(\boldsymbol{x}_0, \boldsymbol{y}_0)||$ is sufficiently small, there exists a solution $\boldsymbol{u}(t)$ of equation 3.3 such that

$$\begin{aligned} \boldsymbol{x}(t) &= \boldsymbol{u}(t) + O(e^{-\gamma t}), \\ \boldsymbol{y}(t) &= \boldsymbol{h}(\boldsymbol{u}(t)) + O(e^{-\gamma t}), \end{aligned}$$

as $t \to \infty$, where γ is a positive constant.

Proposition 3 asserts that the parameter $(\boldsymbol{x}, \boldsymbol{y})$ approaches the center manifold $\boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x})$ quickly and then evolves along it. Thus, the dynamical system, equation 3.2, around the origin is essentially controlled by the slow parameter $\boldsymbol{x}$ and reduced to the lower-dimensional system.

3.2 Main Results. Let us return to the analysis of an (n-2-1)-perceptron. In a column vector representation, the dynamical system, equation 1.5, for the (n-2-1)-perceptron is written as

$$\dot{\boldsymbol{\theta}} = -\left(\frac{\partial L_{(2)}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta})\right)^T.$$

We consider another coordinate system $\boldsymbol{\xi} = \boldsymbol{\xi}(\boldsymbol{\theta})$ and investigate the dynamical system in it. By the coordinate transformation, the dynamical system above is transformed to

$$\dot{\boldsymbol{\xi}} = -\frac{\partial \boldsymbol{\xi}}{\partial \boldsymbol{\theta}} \left(\frac{\partial \boldsymbol{\xi}}{\partial \boldsymbol{\theta}} \right)^T \left(\frac{\partial L_{(2)}}{\partial \boldsymbol{\xi}}(\boldsymbol{\xi}) \right)^T. \tag{3.4}$$

Thus, the coefficient matrix of its linearization at a critical point $\boldsymbol{\xi} = \boldsymbol{\xi}^*$ is

$$\begin{aligned} \frac{\partial \dot{\boldsymbol{\xi}}}{\partial \boldsymbol{\xi}}(\boldsymbol{\xi}^*) &= -\frac{\partial}{\partial \boldsymbol{\xi}} \left\{ \frac{\partial \boldsymbol{\xi}}{\partial \boldsymbol{\theta}} \left(\frac{\partial \boldsymbol{\xi}}{\partial \boldsymbol{\theta}} \right)^T \left(\frac{\partial L_{(2)}}{\partial \boldsymbol{\xi}}(\boldsymbol{\xi}) \right)^T \right\} \Bigg|_{\boldsymbol{\xi}=\boldsymbol{\xi}^*} \\ &= -\frac{\partial \boldsymbol{\xi}}{\partial \boldsymbol{\theta}} \left(\frac{\partial \boldsymbol{\xi}}{\partial \boldsymbol{\theta}} \right)^T \frac{\partial^2 L_{(2)}}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}}(\boldsymbol{\xi}^*), \end{aligned}$$

where we used $(\partial L_{(2)}/\partial \boldsymbol{\xi})(\boldsymbol{\xi}^*) = \mathbf{0}$. This relation implies that the coefficient matrix has the same rank as the Hessian matrix $(\partial^2 L_{(2)}/\partial \boldsymbol{\xi} \partial \boldsymbol{\xi})(\boldsymbol{\xi}^*)$. In particular, the rank of the coefficient matrix of the linearization does not depend on the choice of a coordinate system.

We focus on the dynamics of the learning process around the two points, $\boldsymbol{\theta} = \boldsymbol{\theta}_0, \boldsymbol{\theta}_1$, which are boundaries of the repulsive and attractive parts of a Milnor-like attractor $\{\boldsymbol{\theta}_\lambda | \lambda \in \mathbb{R}\}$. Concretely, they are denoted as $\boldsymbol{\theta}_0 = (\boldsymbol{w}^*, \boldsymbol{w}^*, 0, v^*)$ and $\boldsymbol{\theta}_1 = (\boldsymbol{w}^*, \boldsymbol{w}^*, v^*, 0)$, where $(\boldsymbol{w}^*, v^*)$ is a minimizer of the loss $L_{(1)}$ for the (n-1-1)-perceptron as mentioned in proposition 1. This is because the rank of the Hessian matrix at $\boldsymbol{\theta}_\lambda$ degenerates by one dimension for $\lambda \neq 0, 1$ and by $n + 2$ dimension for $\lambda = 0, 1$, which is shown in appendix A.

We introduce a new coordinate system $\boldsymbol{\xi} = (\boldsymbol{w}, v, \boldsymbol{u}, z)$ by

$$\begin{cases} \boldsymbol{w} = \dfrac{v_1 (\boldsymbol{w}_1 - \boldsymbol{w}^*) + v_2 (\boldsymbol{w}_2 - \boldsymbol{w}^*)}{v^*} + \boldsymbol{w}^* \\ v = v_1 + v_2 \\ \boldsymbol{u} = \dfrac{v_2 (\boldsymbol{w}_1 - \boldsymbol{w}^*) - v_1 (\boldsymbol{w}_2 - \boldsymbol{w}^*)}{v^*} \\ z = v_1 - v_2 \end{cases} \tag{3.5}$$

This formula defines a coordinate system on the region $\{v_1^2 + v_2^2 \neq 0\}$. Under the coordinate system, equation 3.5, the critical points $\boldsymbol{\theta}_\lambda$ are denoted as

$$\boldsymbol{\xi}_\lambda = (\boldsymbol{w}^*, v^*, \mathbf{0}, (2\lambda - 1)v^*).$$

In particular, $\boldsymbol{\xi}_0 = (\boldsymbol{w}^*, v^*, \mathbf{0}, -v^*)$ and $\boldsymbol{\xi}_1 = (\boldsymbol{w}^*, v^*, \mathbf{0}, v^*)$. Now we arrive at the main theorem of this article:

Theorem 2. *In the coordinate system $\boldsymbol{\xi} = (\boldsymbol{w}, v, \boldsymbol{u}, z)$, the dynamical system, equation 1.5, admits a center manifold structure around the critical points $\boldsymbol{\theta} = \boldsymbol{\theta}_0, \boldsymbol{\theta}_1$ in which $(\boldsymbol{w}, v)$ converge exponentially fast.*

To prove the theorem, we make use of the following lemma.

Lemma 2. *If the matrix X is positive definite and Y is positive semidefinite, all the eigenvalues of the matrix XY are nonnegative.*

Proof. The matrix XY is rewritten as

$$XY = X^{\frac{1}{2}}(X^{\frac{1}{2}} Y X^{\frac{1}{2}})X^{-\frac{1}{2}},$$

where $X^{\frac{1}{2}}$ is a unique positive-definite matrix such that $(X^{\frac{1}{2}})^2 = X$. Here, the matrix $Z := X^{\frac{1}{2}} Y X^{\frac{1}{2}}$ is positive semidefinite. Hence, for each eigenvector $\boldsymbol{a}$ of Z, the vector $X^{\frac{1}{2}}\boldsymbol{a}$ is an eigenvector of the matrix XY, and the corresponding eigenvalue is nonnegative. □

Proof of Theorem 2. The proof is essentially based on a straightforward calculation. The coefficient matrix of the linearization of the dynamical system, equation 3.4, under the coordinate system, equation 3.5, splits into $(\boldsymbol{w}, v)$ part and $(\boldsymbol{u}, z)$ part for $\lambda = 0, 1$. In fact, for $\lambda \in \mathbb{R}$, the negative of the coefficient matrix is written as

$$\tilde{A}_\lambda = \begin{array}{c} \\ \boldsymbol{w}, v \\ \boldsymbol{u} \\ z \end{array} \overbrace{\phantom{\begin{pmatrix} a & b \\ c & d \end{pmatrix}}}^{\boldsymbol{w}, v} \overbrace{\phantom{\begin{pmatrix} a \\ c \end{pmatrix}}}^{\boldsymbol{u}} \overbrace{\phantom{\begin{pmatrix} a \\ c \end{pmatrix}}}^{z} \begin{pmatrix} (1+2k_\lambda)Q + \frac{1+3k_\lambda}{1+2k_\lambda}H & (1+2k_\lambda)P & -\frac{(-1+2\lambda)k_\lambda}{1+2k_\lambda}H & 0 \\ 2P^T & 2R & 0 & 0 \\ -\frac{(-1+2\lambda)k_\lambda}{1+2k_\lambda}H & 0 & -\frac{k_\lambda}{1+2k_\lambda}H & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

and the system is written as

$$\dot{\boldsymbol{\xi}} = -\tilde{A}_\lambda(\boldsymbol{\xi} - \boldsymbol{\xi}_\lambda) + \tilde{\boldsymbol{g}}_\lambda(\boldsymbol{\xi}),$$

where $\tilde{\boldsymbol{g}}_\lambda$ is the higher-order term, which vanish at the $\boldsymbol{\xi} = \boldsymbol{\xi}_\lambda$ together with its first derivative. Here,

$$\begin{aligned} k_\lambda &:= (1-\lambda)\lambda, \\ P &:= \mathbb{E}_x\left[(\partial^2\ell)\, v^*\, \varphi(\boldsymbol{w}^* \cdot \boldsymbol{x})\, \varphi'(\boldsymbol{w}^* \cdot \boldsymbol{x})\, \boldsymbol{x}\right], \\ Q &:= \mathbb{E}_x\left[(\partial^2\ell)\, (v^*)^2\, \varphi'(\boldsymbol{w}^* \cdot \boldsymbol{x})^2\, \boldsymbol{x}\boldsymbol{x}^T\right], \\ R &:= \mathbb{E}_x\left[(\partial^2\ell)\, \varphi(\boldsymbol{w}^* \cdot \boldsymbol{x})^2\right], \end{aligned}$$

$$\partial \ell := \frac{\partial \ell(\boldsymbol{x}, f_{(1)}(\boldsymbol{x}; \boldsymbol{\theta}^*))}{\partial y},$$
$$\partial^2 \ell := \frac{\partial^2 \ell(\boldsymbol{x}, f_{(1)}(\boldsymbol{x}; \boldsymbol{\theta}^*))}{\partial y^2},$$

and H is the matrix defined by equation 2.1. H and Q are matrices, P is a column vector, and R is a scalar. For $\lambda = 0$ or 1, the negative of the coefficient matrix is of the form

$$\tilde{A}_0 = \tilde{A}_1 = \begin{array}{r} \\ \boldsymbol{w}, \upsilon \\ \boldsymbol{u} \\ z \end{array} \overbrace{\begin{array}{cc} \boldsymbol{w}, \upsilon & \end{array}}^{} \overbrace{\begin{array}{c} \boldsymbol{u} \end{array}}^{} \overbrace{\begin{array}{c} z \end{array}}^{} \begin{pmatrix} Q+H & P & 0 & 0 \\ 2P^T & 2R & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{3.6}$$

We show that all the eigenvalues of $(\boldsymbol{w}, \upsilon)$-block of the coefficient matrix $-\tilde{A}_0$ are strictly negative. Recall that the coefficient matrix at a critical point $\boldsymbol{\xi}^*$ of the dynamical system, equation 1.5, is given by

$$-\frac{\partial \boldsymbol{\xi}}{\partial \boldsymbol{\theta}} \left(\frac{\partial \boldsymbol{\xi}}{\partial \boldsymbol{\theta}} \right)^T \frac{\partial^2 L_{(2)}}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}} (\boldsymbol{\xi}^*).$$

Applying lemma 2 to $X = (\partial \boldsymbol{\xi}/\partial \boldsymbol{\theta})(\partial \boldsymbol{\xi}/\partial \boldsymbol{\theta})^T$ and $Y = (\partial^2 L_{(2)}/\partial \boldsymbol{\xi} \partial \boldsymbol{\xi})(\boldsymbol{\xi}_0)$, all the eigenvalues of the coefficient matrix $-\tilde{A}_0$ are nonpositive. One can see that the Hessian matrix is positive semidefinite and degenerates by $n+2$ dimension at $\boldsymbol{\theta} = \boldsymbol{\theta}_0, \boldsymbol{\theta}_1$ in appendix A. Since a coordinate transformation preserves the rank of the coefficient matrix of linearization, $\tilde{A}_0$ degenerates by $n+2$ dimension, which is equal to the size of $(\boldsymbol{u}, z)$-block. This implies that the $(\boldsymbol{w}, \upsilon)$-block is of full rank, and thus all the eigenvalues of $(\boldsymbol{w}, \upsilon)$-block are strictly negative. It is proved similarly for $\boldsymbol{\xi} = \boldsymbol{\xi}_1$.

Due to proposition 2, there are center manifolds parametrized by $(\boldsymbol{u}, z)$ around $\boldsymbol{\theta} = \boldsymbol{\theta}_0, \boldsymbol{\theta}_1$ respectively. □

3.3 Reduced Dynamical System. By virtue of proposition 3 and theorem 2, we can assume that the dynamics of the gradient descent is on the center manifold near the points $\boldsymbol{\theta} = \boldsymbol{\theta}_0, \boldsymbol{\theta}_1$. Thus, we can reduce the dynamical system into that of $(\boldsymbol{u}, z)$. Recalling the coefficient matrix, equation 3.6, we can see that $\dot{\boldsymbol{u}}$ and $\dot{z}$ have no first-order terms. In more detail, calculating the Taylor expansion of $(\dot{\boldsymbol{u}}, \dot{z})$ up to the second order around $\boldsymbol{\xi} = \boldsymbol{\xi}_1$, we obtain

$$\dot{\boldsymbol{u}} = \frac{1}{v^*}\Big\{-(P\cdot(\boldsymbol{w}-\boldsymbol{w}^*))(\boldsymbol{u}+(\boldsymbol{w}-\boldsymbol{w}^*)) - (v-v^*)(RI+\frac{1}{2}H)(\boldsymbol{u}+(\boldsymbol{w}-\boldsymbol{w}^*)) + \frac{1}{2}(z-v^*)H(\boldsymbol{u}+(\boldsymbol{w}-\boldsymbol{w}^*))\Big\} + O(||\boldsymbol{\xi}-\boldsymbol{\xi}_1||^3), \tag{3.7}$$

$$\dot{z} = \frac{1}{v^*}\Big\{-(\boldsymbol{w}-\boldsymbol{w}^*)^T Q(\boldsymbol{u}+(\boldsymbol{w}-\boldsymbol{w}^*)) - (v-v^*)(P\cdot(\boldsymbol{u}+(\boldsymbol{w}-\boldsymbol{w}^*))) - \frac{1}{2}(\boldsymbol{w}-\boldsymbol{w}^*)^T H(\boldsymbol{w}-\boldsymbol{w}^*) + \frac{1}{2}\boldsymbol{u}^T H\boldsymbol{u}\Big\} + O(||\boldsymbol{\xi}-\boldsymbol{\xi}_1||^3), \tag{3.8}$$

where I denotes the $(n+1)\times(n+1)$ identity matrix. Now we consider the reduced dynamical system on the center manifold. Here, the center manifold $(\boldsymbol{w}, v) = \boldsymbol{h}(\boldsymbol{u}, z)$ satisfies that

$$\boldsymbol{h}(\boldsymbol{u},z) = \begin{bmatrix} \boldsymbol{w}(\boldsymbol{u},z) \\ v(\boldsymbol{u},z)\end{bmatrix} = \begin{bmatrix}\boldsymbol{w}^* \\ v^*\end{bmatrix} + O(\|\boldsymbol{u}, z-v^*\|^2),$$

by definition. This gives an approximation of the dynamics on the center manifold near $\boldsymbol{\xi} = \boldsymbol{\xi}_1$ as

$$\dot{\boldsymbol{u}} = \frac{1}{2v^*}(z-v^*)H\boldsymbol{u} + O(\|\boldsymbol{u}, z-v^*\|^3),$$
$$\dot{z} = \frac{1}{2v^*}\boldsymbol{u}^T H\boldsymbol{u} + O(\|\boldsymbol{u}, z-v^*\|^3). \tag{3.9}$$

Neglecting the higher-order terms, we can integrate this equation to obtain

$$\|\boldsymbol{u}\|^2 = (z-v^*)^2 + C, \tag{3.10}$$

where C is an integral constant.

Around the point $\boldsymbol{\xi} = \boldsymbol{\xi}_0$, we obtain the similar dynamics,

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

and the relation,

$$\|\boldsymbol{u}\|^2 = (z+v^*)^2 + C.$$

We remark that theorem 2 is valid even when there exists a true parameter in the singular region $R(\boldsymbol{w}^*, v^*)$; however, in this case, such a simple form of the reduced dynamical system as equation 3.9 is not obtained. As mentioned above, this case implies that H becomes the zero matrix. Then, the second-order terms of the reduced dynamical system, equation 3.9, vanish, and the third-order terms become dominant. Thus, we have to take into account the cross terms between $(\boldsymbol{w} - \boldsymbol{w}^*, v - v^*)$ and $(\boldsymbol{u}, z - v^*)$. It needs to calculate the center manifold $(\boldsymbol{w}, v) = \boldsymbol{h}(\boldsymbol{u}, z)$ up to the second order, which makes the analysis complicated.

Finally, we remark on a difference between our analysis and previous work. Wei et al. (2008) have studied a reduced dynamical system in the vicinity of the whole part of a Milnor-like attractor. On the other hand, a center manifold is defined locally, and center manifolds around each of two points cannot be connected at a midpoint in general. Thus, one cannot discuss a center manifold defined around the entire region of a Milnor-like attractor.

3.4 More General Models. Our results can be extended to a more general model including multilayer perceptrons whose output is one-dimensional. In this section, we consider a parameterized family of functions that can be written as

$$\begin{aligned} f(\boldsymbol{x}; \boldsymbol{\theta}) &:= g(\boldsymbol{x}, \boldsymbol{\tau}) + v_1 \phi(\boldsymbol{x}; \boldsymbol{w}_1, \boldsymbol{\tau}) + v_2 \phi(\boldsymbol{x}; \boldsymbol{w}_2, \boldsymbol{\tau}), \\ \boldsymbol{\theta} &= (\boldsymbol{w}_1, \boldsymbol{w}_2, v_1, v_2, \boldsymbol{\tau}), \end{aligned} \tag{3.11}$$

where we assume that $g(\boldsymbol{x}; \boldsymbol{\tau})$ and $\phi(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{\tau})$ are twice differentiable with respect to $\boldsymbol{\tau}$ and $(\boldsymbol{w}, \boldsymbol{\tau})$, respectively.

A multilayer perceptron $\tilde{\alpha}^{(L)}(\boldsymbol{x}; \boldsymbol{\theta})$ with L layers defined recursively as

$$\begin{aligned} \alpha^{(0)}(\boldsymbol{x}; \boldsymbol{\theta}) &:= (1, x_1, \ldots, x_{n_0})^T, \\ \tilde{\alpha}^{(\ell)}(\boldsymbol{x}; \boldsymbol{\theta}) &:= W^{(\ell)} \alpha^{(\ell-1)}(\boldsymbol{x}; \boldsymbol{\theta}), \\ \alpha^{(\ell)}(\boldsymbol{x}; \boldsymbol{\theta}) &:= \left(1, \varphi(\tilde{\alpha}_1^{(\ell)}(\boldsymbol{x}; \boldsymbol{\theta})), \ldots, \varphi(\tilde{\alpha}_{n_\ell}^{(\ell)}(\boldsymbol{x}; \boldsymbol{\theta}))\right)^T, \quad 1 \le \ell \le L, \\ \boldsymbol{\theta} &= (W^{(1)}, \ldots, W^{(L)}), \end{aligned}$$

for each $\boldsymbol{x} = (x_1, \ldots, x_{n_0})^T \in \mathbb{R}^{n_0}$, where $W^{(\ell)}$ is an $n_\ell \times (n_{\ell-1} + 1)$ matrix, $n_0, n_1, \ldots, n_L \in \mathbb{N}$, and φ is a twice-differentiable activation function. Assuming that the dimension n_L of the output is equal to 1 and denoting

$$\begin{aligned} W^{(L-1)} &= \begin{bmatrix} \boldsymbol{w}_1 & \boldsymbol{w}_2 & \cdots & \boldsymbol{w}_{n_{L-1}} \end{bmatrix}^T, \\ W^{(L)} &= \begin{bmatrix} v_0 & v_1 & v_2 & \cdots & v_{n_{L-1}} \end{bmatrix}, \end{aligned}$$

a multilayer perceptron is represented as model 3.11 by letting

$$\phi(x; w, \tau) := \varphi(w \cdot \alpha^{(L-2)}(x; \tau)),$$
$$g(x; \tau) := \sum_{i=3}^{n_{L-1}} v_i \varphi(w_i \cdot \alpha^{(L-2)}(x; \tau)),$$
$$\tau := (W^{(1)}, \ldots, W^{(L-2)}, w_3, \ldots, w_{n_{L-1}}, v_0, v_3, \ldots, v_{n_{L-1}}).$$

Our main result is extended to the model, equation 3.11 as follows. Let $\theta^* = (w^*, v^*, \tau^*)$ be a strict local minimizer of an averaged loss function for the degenerate model:

$$f_{(1)}(x; \tilde{w}, \tilde{v}, \tau) = g(x; \tau) + \tilde{v}\, \phi(x; \tilde{w}, \tau).$$

Then the coordinate system $\xi = (w, v, \tau, u, z)$, given by formula 3.5, admits a center manifold structure around the two points $\theta = (w^*, w^*, 0, v^*, \tau^*)$, $(w^*, w^*, v^*, 0, \tau^*)$, and the dynamical system is reduced to that of (u, z). This is confirmed by the argument of the coefficient matrix of the linearization, similar to theorem 2.

4 Numerical Simulations

In the previous section, we showed that the dynamics of (w, v) are fast and those of (u, z) are slow under the coordinate system, equation 3.5. In this section, we verify this fact by numerical simulations.

4.1 Example 1. As the first example, we set the input dimension to be $n = 1$ and choose the teacher function $T : \mathbb{R} \to \mathbb{R}$ defined by

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

where tanh is the hyperbolic tangent function. The shape of T is shown in Figure 2 by the solid black line. 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:

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

and the true parameter is $(w_1, w_2, v_1, v_2) = (1, 4, 2, -1)$. We also discard the bias terms w_1^0 and w_2^0 of the student perceptron. This makes the matrix H scalar valued, and hence the assumption of proposition 1 holds trivially.

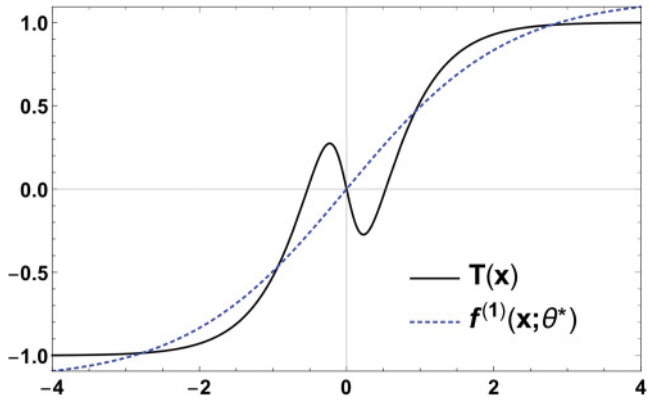


Figure 2: The target function $T(x)$ and the (1-1-1)-perceptron $f_{(1)}(x;\boldsymbol{\theta}^*)$, which corresponds to the local minimizer $\boldsymbol{\theta}^*$.

We assume that the data set $\{x_s\}_{s=1}^S$ is given and that the probability distribution of x is the empirical distribution on that data set. Then the transition formula, equation 1.4, of the parameter $\boldsymbol{\theta}$ is rewritten as

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

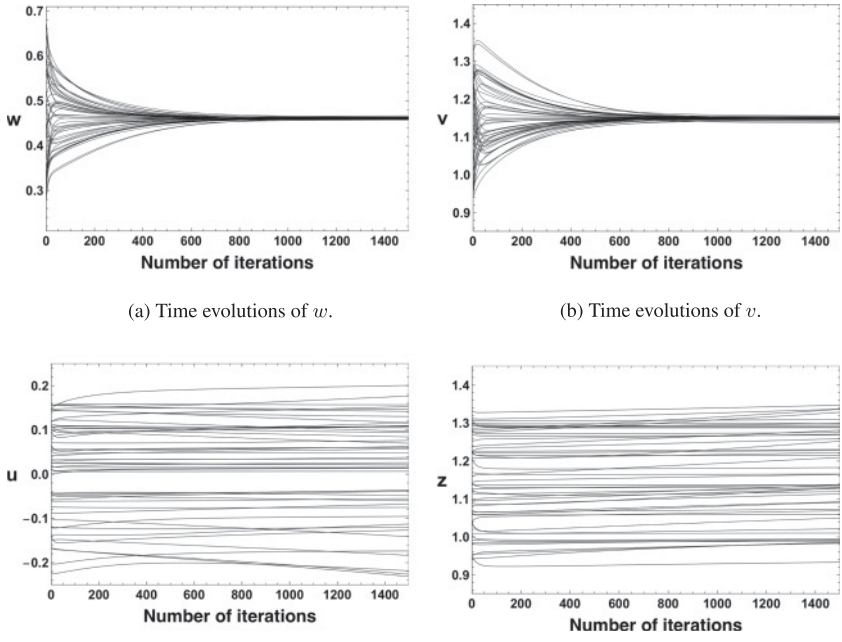
In this simulation, we set the size S of the data set to be 1000, and data $\{x_s\}_{s=1}^S$ are drawn identical and independenty distributed (i.i.d.) according to $N(0, 2^2)$. Here, $N(\mu, \sigma^2)$ denotes the gaussian distribution with mean μ and variance σ^2.

For a data set given as above, we obtained a local minimizer $\boldsymbol{\theta}^* = (w^*, v^*) = (0.459, 1.15)$ of $L_{(1)}$. The shape of the function that corresponds to the local minimizer is shown by the dashed blue line in Figure 2. The value of H is approximately 0.0472. Since $H > 0$, the attractive region is $\{\boldsymbol{\theta}_\lambda \mid \lambda \in (0, 1)\}$, due to proposition 1.

Figure 3 displays time evolutions of each parameter in the first 1500 iterations from 50 different initial points. We chose an initial parameter $\boldsymbol{\theta}^{(0)} = (w_1^{(0)}, w_2^{(0)}, v_1^{(0)}, v_2^{(0)})$ 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) 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 for the first 1500 iterations. Each trajectory of w or v quickly converges to the equilibrium point $w^* = 0.459$, $v^* = 1.15$, respectively. Trajectories of u and z evolve very slowly compared with w and v.

$[a, b] \subset \mathbb{R}$. We set the learning rate ε to be 0.05 and the number of iterations to be 20,000. We can see that the parameters w and v converge to their equilibriums exponentially fast (see Figures 3a and 3b), while u and z evolve slowly (see Figures 3c and 3d).

Figure 4 shows evolutions on the $(z, ||u||^2)$-plane. The red circles in the figure represent initial points. When $w = w^*$ and $v = v^*$, the z-axis is a Milnor-like attractor, and the region $|z| < v^*$ is the attractive part of it. We can check that parameters near the attractive region are trapped and those near the repulsive region are escaping. The intersection point of the line $z = v^*$ and z-axis corresponds to the point $\boldsymbol{\theta} = \boldsymbol{\theta}_1$, the boundary of the attractive and repulsive parts of the Milnor-like attractor. The analytical trajectories, equation 3.10, are plotted as dashed blue curves. Numerical evolutions of the parameter follow the analytical trajectories considerably well around $\boldsymbol{\theta} = \boldsymbol{\theta}_1$. We see in the figure that some instances of time evolutions change their direction sharply. This is because the fast dynamics of w and v are the main dynamics at the beginning of the learning, while the

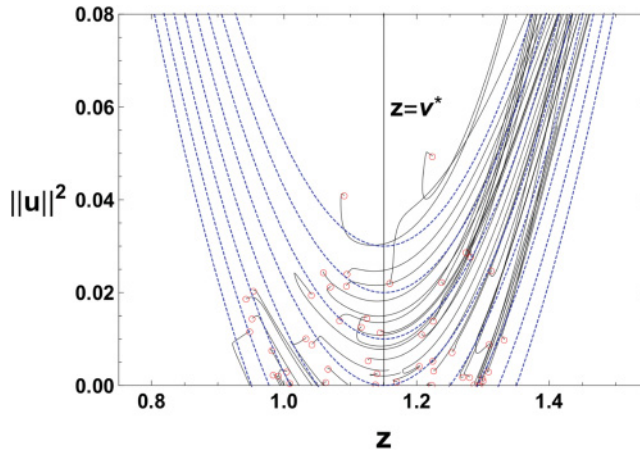


Figure 4: Trajectories on the $(z, ||u||^2)$-plane obtained by learning for 20,000 iterations (solid black curves) and analytical trajectories (dashed blue curves) near $\boldsymbol{\theta} = \boldsymbol{\theta}_1 = (w^*, w^*, v^*, 0)$. Red circles represent initial points.

slow dynamics of u and z become dominant after w and v converge to the center manifold.

4.2 Example 2. As the second example, we consider a three-layer perceptron whose input dimension is $n = 2$. Let the teacher function $T : \mathbb{R}^2 \to \mathbb{R}$ be given by

$$T(x_1, x_2) := 0.75\,\mathrm{Sgm}(2.5x_1 - 2.5x_2) + \mathrm{Sgm}(2.5x_1 + 2.5x_2) + 0.5,$$

where Sgm is the logistic sigmoidal function, which is defined by

$$\mathrm{Sgm}(x) := \frac{1}{1 + e^{-2x}} = \frac{1}{2}(1 + \tanh(x)).$$

Figure 5 shows the shape of the teacher function $T(\boldsymbol{x})$. We use a perceptron with no bias terms also in this simulation and choose Sgm as the activation function. Note that a (2-2-1)-perceptron is unable to represent the target function in this case.

Also in this simulation, we assume that the probability distribution of $\boldsymbol{x}$ is the empirical distribution on a fixed data set $\{\boldsymbol{x}_s\}_{s=1}^S$. We set the number S of the data set to be 1000, and draw a data set $\{\boldsymbol{x}_s\}_{s=1}^S$ i.i.d. according to $N(\mathbf{0}, I_2)$, where I_2 denotes the 2×2 identity matrix. We chose a realization $\{\boldsymbol{x}_s\}_{s=1}^S$ as above and obtained a local minimizer $\boldsymbol{\theta}^* = (\boldsymbol{w}^*, v^*)$ of $L_{(1)}$, where $\boldsymbol{w}^* = (0.399, 0.0652)$ and $v^* = 2.76$. Figure 6 shows the shape of the

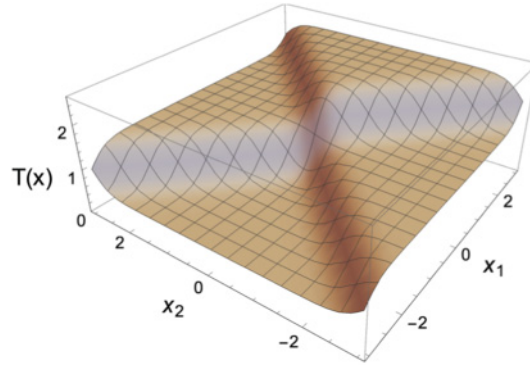


Figure 5: The target teacher function $T(\boldsymbol{x})$.

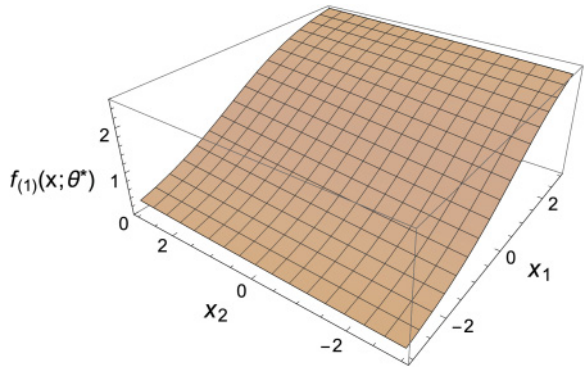


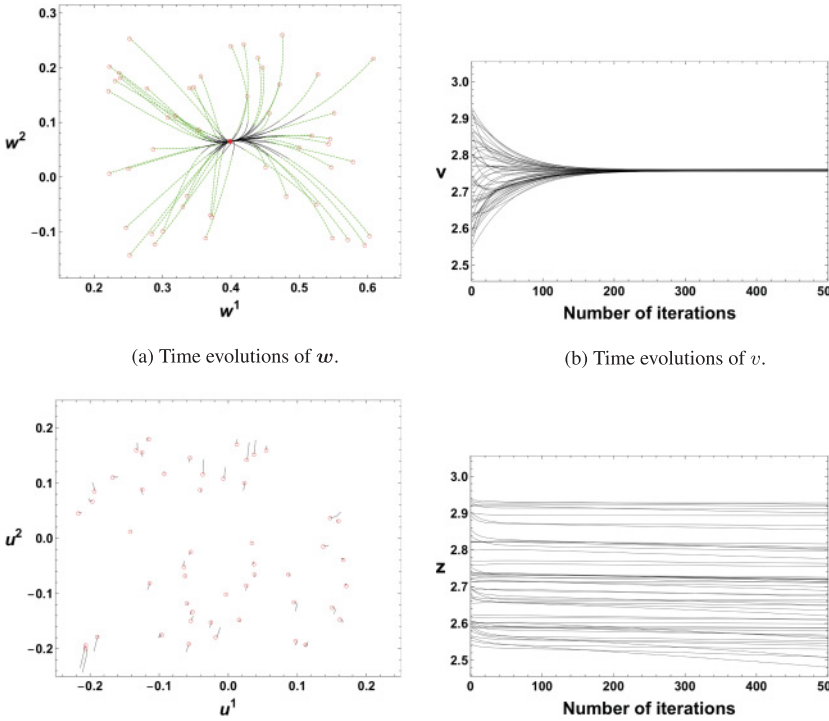
Figure 6: The (2-1-1)-perceptron $f_{(1)}(\boldsymbol{x};\boldsymbol{\theta}^*)$ that corresponds to the local minimizer $\boldsymbol{\theta}^*$.

(2-1-1)-perceptron corresponding to the local minimizer. The matrix H is numerically computed as

$$\begin{pmatrix} -0.044 & -0.026 \\ -0.026 & -0.20 \end{pmatrix}.$$

Since this H is negative definite, the attractive region is $\{\boldsymbol{\theta}_\lambda \mid \lambda \in \mathbb{R} \setminus [0,1]\}$ due to proposition 1.

Figures 7a to 7d show time evolutions of each parameter in the first 500 iterations from 50 different initial points. We chose initial parameters of the (2-2-1)-perceptron as



(a) Time evolutions of $\boldsymbol{w}$.

(b) Time evolutions of v.

(c) Time evolutions of $\boldsymbol{u}$.

(d) Time evolutions of z.

Figure 7: Time evolutions of each parameter for the first 500 iterations. Each trajectory of w or v quickly converges to the equilibrium point $\boldsymbol{w}^* = (0.399, 0.0652)$, $v^* = 2.76$ respectively. However, trajectories of u and z evolve very slowly compared with w and v. In panel a, the first 30 iterations are dashed in green to display the speed of convergence, and the red point at the center represents $\boldsymbol{w} = \boldsymbol{w}^*$. Red circles in panels a and c represent initial points.

$$\boldsymbol{w}_1^{(0)} = \boldsymbol{w}^* + \boldsymbol{\zeta}_1, \quad \boldsymbol{w}_2^{(0)} = \boldsymbol{w}^* + \boldsymbol{\zeta}_2,$$
$$v_1^{(0)} = v^* + \frac{1}{2}(\zeta_3 + \zeta_4), \quad v_2^{(0)} = \frac{1}{2}(\zeta_3 - \zeta_4),$$

where $\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2 \sim U(-0.2, 0.2)^2$ and $\zeta_3, \zeta_4 \sim U(-0.2, 0.2)$. We set the learning rate ε to be 0.05 and the number of iterations to be 20,000. In this simulation, since $\boldsymbol{w}$ and $\boldsymbol{u}$ are two-dimensional, their evolutions are not displayed as time series but as trajectories on each plane. The red circles in Figures 7a and 7c represent initial values of $\boldsymbol{w}$ and $\boldsymbol{u}$, respectively. Figures 7a and 7b show that the parameters $\boldsymbol{w}$ and v converge to their equilibrium $\boldsymbol{w}^*$ and v^* very quickly. To display how quick the convergence is, the first 30 iterations

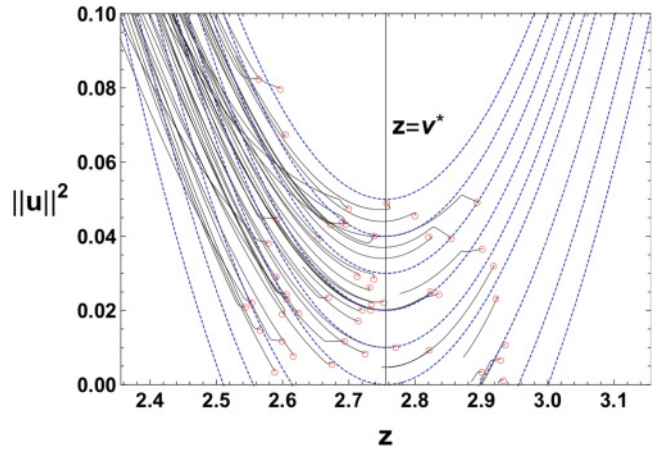


Figure 8: Trajectories on the $(z, ||\boldsymbol{u}||^2)$-plane obtained by learning for 20,000 iterations (solid black curves) and analytical trajectories near $\boldsymbol{\theta} = \boldsymbol{\theta}_1$ (dashed blue curves). Red circles represent the initial points.

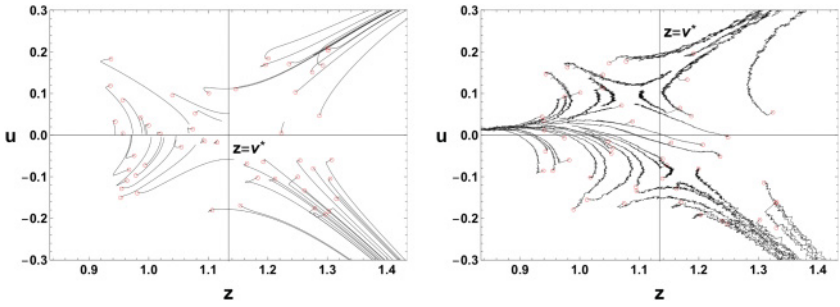
are dashed in green in Figure 7a. In contrast, Figures 7c and 7d show that the parameters $\boldsymbol{u}$ and z evolve very slowly.

Figure 8 plots time evolutions of the parameter $\boldsymbol{\theta}$ on the $(z, ||\boldsymbol{u}||^2)$-plane, which means the plane whose axes indicate the values of z and $\|u\|^2$. We can check that parameters near the attractive part $\{|z| > v^*, ||\boldsymbol{u}||^2 = 0\}$ of the Milnor-like attractor are trapped and that those near the repulsive part $\{|z| < v^*, ||\boldsymbol{u}||^2 = 0\}$ are escaping. The numerical evolutions follow the analytical flows (dashed blue curves) well also in this case.

5 Aspects of Online Learning

In this section, we discuss stochastic effects in the learning process. Thus far, we have analyzed the dynamical system, equation 1.5, driven by the averaged gradient. In practice, the averaged gradient is estimated by the arithmetic mean of the instantaneous gradient $\partial \ell(\boldsymbol{x}, f_{(d)}(\boldsymbol{x}; \boldsymbol{\theta}))/\partial \boldsymbol{\theta}$ over a large number of input data. However, taking the arithmetic mean for each update of the parameter demands high computational cost. In order to reduce the cost, the expectation is often replaced by a single realization of the instantaneous gradient. Such a method is called *online learning*, a typical stochastic gradient descent method. Mathematically, it is given by

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \varepsilon \left. \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{x}_t, f_{(d)}(\boldsymbol{x}_t; \boldsymbol{\theta})) \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}, \tag{5.1}$$



(a) Averaged gradient descent method.

(b) On-line learning.

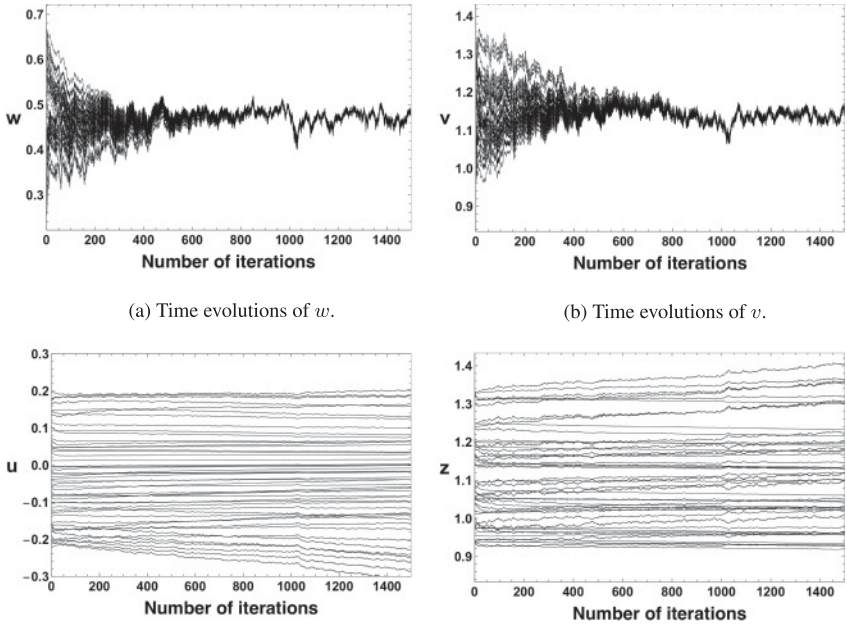
Figure 9: Trajectories on the (z, u)-plane obtained by the averaged gradient descent method, equation 1.5, and the online learning, equation 5.1, for 20,000 iterations. Red circles represent initial values.

where $\{\boldsymbol{x}_t\}_t$ are i.i.d. realizations of the input data $\boldsymbol{x}$. Unlike the deterministic dynamical system, equation 1.5, the system 5.1 is a randomized dynamical system.

In numerical simulations, we found that sample paths of the online learning seem quite different from trajectories obtained in the averaged gradient descent method. We set the distribution over which the loss function $L_{(1)}$ is averaged to be $N(0, 2^2)$ and obtained a local minimizer of $L_{(1)}$ as $\boldsymbol{\theta}^* = (w^*, v^*) \approx (0.472, 1.13)$. We carried out numerical simulations of the online learning in the same setting as example 1 in section 4. Figure 9a shows numerical trajectories of the averaged gradient descent on the (z, u)-plane around $\boldsymbol{\theta} = \boldsymbol{\theta}_1$. In order to approximate the averaged gradient descent sufficiently, we used the empirical distribution on a data set of 10,000 data drawn i.i.d. according to $N(0, 2^2)$. Figure 9b shows sample paths of the online learning for a common input data sequence $\{x_t\}_t$. In contrast to the averaged gradient descent, in the online learning, some sample paths move from region $\{|z| > v^*\}$ to $\{|z| < v^*\}$. Such sample paths are observed even when we use another realization of the input data sequence, and its dynamics seems qualitatively different from the averaged one.

In order to investigate this phenomenon, we observe the evolution of the parameters, again in the coordinate system, equation 3.5. Figures 10a to 10d show time evolutions of each parameter in the first 1500 iterations of the online learning. The parameters (w, v) evolve very fast compared with (u, z) also in this case. However, in this case, (w, v) does not converge to its equilibrium point $(w^*, v^*) \approx (0.472, 1.13)$, but fluctuates stochastically around (w^*, v^*).

Based on these observations, we suppose that w and v run over a sufficiently wide range of their values to be integrated, while u and z move in a small range. Then we assume that the dynamics of (u, z) is integrated



(a) Time evolutions of w.

(b) Time evolutions of v.

(c) Time evolutions of u.

(d) Time evolutions of z.

Figure 10: Time evolutions of each parameter for the first 1500 iterations of the online learning. Each trajectory of w or v fluctuates intensively around the equilibrium point $w^* \approx 0.472$, $v^* \approx 1.134$, respectively. Trajectories of u and z evolve very slowly compared with w and v in this case.

with respect to (w, v) according to some stationary distribution. We further assume that (w, v) are distributed around (w^*, v^*) with finite variance. By integrating the Taylor expansions, equations 3.7 and 3.8, with (w, v), we obtain the following dynamical system near $\boldsymbol{\theta} = \boldsymbol{\theta}_1$:

$$\dot{\boldsymbol{u}} = \frac{1}{2}(z - v^*)H\boldsymbol{u} + C_1,$$
$$\dot{z} = \frac{1}{2}\boldsymbol{u}^T H\boldsymbol{u} + C_2. \tag{5.2}$$

Here, C_1 and C_2 are constants resulting from the variance and covariance of (w, v). Figure 11 shows the analytical trajectories of the dynamical system, equation 5.2, where $C_1 = 1.71 \times 10^{-4}$ and $C_2 = -3.06 \times 10^{-4}$ are determined heuristically. One can find that the deterministic dynamical system, equation 5.2, gives similar trajectories to sample paths of the online learning presented in Figure 9b.

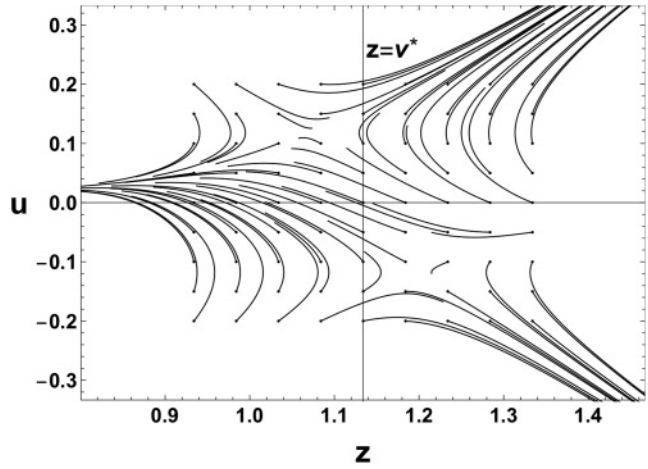


Figure 11: Analytical trajectories on the (z, u)-plane given by the dynamical system 5.2 with $C_1 = 1.71 \times 10^{-4}$ and $C_2 = -3.06 \times 10^{-4}$.

We deduce that a fluctuation of the parameter around a center manifold causes constants C_1 and C_2 working as drift terms and that it makes the dynamics of the online learning qualitatively different from those of the averaged gradient descent. This example suggests that stochastic effects can influence a macroscopic flow of the learning process via a center manifold structure.

6 Conclusion

In this article, we first gave a quick review of a mechanism that causes plateau phenomena in a three-layer perceptron—in particular, how degeneration of hidden units gives rise to a Milnor-like attractor consisting of both attractive and repulsive parts. We next investigated the dynamics of learning around special points on a Milnor-like attractor and proved the existence of the center manifold. We also succeeded in integrating the reduced dynamical system to obtain an analytical form of a trajectory. We performed several numerical simulations to demonstrate the accuracy of our results. As an application of the center manifold structure, we gave an explanation for a characteristic behavior of the online learning.

Unfortunately, the two examples presented in section 4 were the only ones that we could find in which the assumptions of proposition 1 are fulfilled. This might suggest that the appearance of a Milnor-like attractor would be a rather rare situation in a perceptron that has bias terms. In fact, just by replacing the activation function Sgm with tanh in example 2, the matrix H becomes indefinite and the assumption of proposition 1

is violated. Finding more suggestive examples that shed light on the complex behavior of the dynamics of learning is an important subject for future study.

In section 5, we investigated stochastic effects of the online learning from an intermediate viewpoint between fully stochastic and averaged dynamics. We made use of the center manifold of the averaged dynamics and discussed an integration with quickly fluctuating parameters. There have been many reports of qualitative differences between stochastic and deterministic methods; however, there are few general theories for analyzing such dissimilarities. We expect that the intermediate viewpoint in this article can be a clue to clarify stochastic effects in learning.

Appendix: Proofs of Proposition 1 and Theorem 1

This appendix gives proofs of proposition 1 and Theorem 1. The proof of proposition 1 is based on the analysis of the Hessian matrix of $L_{(2)}$. However, the Hessian at the point $\boldsymbol{\theta}_\lambda$ becomes singular, since $L_{(2)}(\boldsymbol{\theta}_\lambda)$ is constant along $\lambda \in \mathbb{R}$. Thus, we need to take into account higher-order derivatives of $L_{(2)}$, which is overlooked in the Fukumizu and Amari (2000). The prototype of theorem 1 was given by Amari et al. (2018); however, they proved it only for a special case. Here, we give a rigorous proof with an additional mild assumption.

Proof of Proposition 1. We introduce a new coordinate system $\boldsymbol{\xi} = (\boldsymbol{w}, v, \boldsymbol{u}, z)$ by

$$\begin{cases} \boldsymbol{w} = \dfrac{v_1\boldsymbol{w}_1 + v_2\boldsymbol{w}_2}{v_1 + v_2} \\ v = v_1 + v_2 \\ \boldsymbol{u} = \boldsymbol{w}_1 - \boldsymbol{w}_2 \\ z = v_1 - v_2 \end{cases}, \tag{A.1}$$

where $v_1 + v_2 \neq 0$. Under this coordinate system, the point $\boldsymbol{\theta}_\lambda$ is denoted as $\boldsymbol{\xi}_\lambda = (\boldsymbol{w}^*, v^*, \mathbf{0}, (2\lambda - 1)v^*)$. Note that each point $\boldsymbol{\xi} = \boldsymbol{\xi}_\lambda$ is a critical point of $L_{(2)}(\boldsymbol{\xi})$ by lemma 1. The inverse transformation is given as

$$\begin{cases} \boldsymbol{w}_1 = \boldsymbol{w} + \dfrac{v - z}{2v}\boldsymbol{u} \\ \boldsymbol{w}_2 = \boldsymbol{w} - \dfrac{v + z}{2v}\boldsymbol{u} \\ v_1 = \dfrac{v + z}{2} \\ v_2 = \dfrac{v - z}{2} \end{cases}$$

We observe the Hessian matrix $\mathrm{Hess}(\boldsymbol{\xi}_\lambda)$ of $L_{(2)}(\boldsymbol{\xi})$ at $\boldsymbol{\xi}_\lambda$ for each $\lambda \in \mathbb{R}$. For all $\boldsymbol{\xi}$ such that $\boldsymbol{u} = \mathbf{0}$ (or, equivalently, $\boldsymbol{w}_1 = \boldsymbol{w}_2$), using the inverse transformation formula above,

$$\begin{aligned}
\frac{\partial L_{(2)}}{\partial \boldsymbol{u}}(\boldsymbol{\xi}) &= \mathbb{E}_x\left[\frac{\partial \ell(\boldsymbol{x}, f_{(2)}(\boldsymbol{x};\boldsymbol{\xi}))}{\partial y}\frac{\upsilon^2 - z^2}{4\upsilon}(\varphi'(\boldsymbol{w}_1\cdot\boldsymbol{x}) - \varphi'(\boldsymbol{w}_2\cdot\boldsymbol{x}))\boldsymbol{x}\right] \\
&= 0, \\
\frac{\partial L_{(2)}}{\partial z}(\boldsymbol{\xi}) &= -\mathbb{E}_x\left[\frac{\partial \ell(\boldsymbol{x}, f_{(2)}(\boldsymbol{x};\boldsymbol{\xi}))}{\partial y}\left(\frac{\upsilon + z}{4\upsilon}\varphi'(\boldsymbol{w}_1\cdot\boldsymbol{x}) - \frac{\upsilon - z}{4\upsilon}\varphi'(\boldsymbol{w}_2\cdot\boldsymbol{x})\right)(\boldsymbol{u}\cdot\boldsymbol{x})\right] \\
&\quad + \mathbb{E}_x\left[\frac{\partial \ell(\boldsymbol{x}, f_{(2)}(\boldsymbol{x};\boldsymbol{\xi}))}{\partial y}\frac{1}{2}(\varphi(\boldsymbol{w}_1\cdot\boldsymbol{x}) - \varphi(\boldsymbol{w}_2\cdot\boldsymbol{x}))\right] \\
&= 0.
\end{aligned}$$

Here, we left $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ for notational simplicity. We then have

$$\begin{aligned}
\frac{\partial^2 L_{(2)}}{\partial \gamma \partial z}(\boldsymbol{\xi}_\lambda) &= 0, \\
\frac{\partial^2 L_{(2)}}{\partial \gamma \partial \boldsymbol{u}}(\boldsymbol{\xi}_\lambda) &= 0,
\end{aligned}$$

where $\gamma = \boldsymbol{w}, \upsilon, z$. Hence, the matrix $\mathrm{Hess}(\boldsymbol{\xi}_\lambda)$ has the form

$$\begin{array}{r} \\ \boldsymbol{w},\upsilon\{ \\ \boldsymbol{u}\{ \\ z\{ \end{array}
\begin{array}{c}
\overbrace{}^{\boldsymbol{w},\upsilon}\ \overbrace{}^{\boldsymbol{u}}\ \overbrace{}^{z} \\
\left(\begin{array}{ccccc}
* & * & 0 & 0 \\
* & * & 0 & 0 \\
0 & 0 & * & 0 \\
0 & 0 & 0 & 0
\end{array}\right).
\end{array}$$

The $(\boldsymbol{w}, \upsilon)$-block is equal to the Hessian matrix of $L_{(1)}$ at $\boldsymbol{\theta}^*$ and is positive definite. In fact, for any $\boldsymbol{\xi}$ such that $\boldsymbol{u} = \mathbf{0}$, noting that $f_{(2)}(\boldsymbol{x};\boldsymbol{\xi}) \equiv f_{(1)}(\boldsymbol{x};\boldsymbol{w},\upsilon)$,

$$\begin{aligned}
\frac{\partial L_{(2)}}{\partial \boldsymbol{w}}(\boldsymbol{\xi}) &= \mathbb{E}_x\left[\frac{\partial \ell(\boldsymbol{x}, f_{(1)}(\boldsymbol{x};\boldsymbol{w},\upsilon))}{\partial y}\upsilon\,\varphi'(\boldsymbol{w}\cdot\boldsymbol{x})\boldsymbol{x}\right] = \frac{\partial L_{(1)}}{\partial \boldsymbol{w}_{(1)}}(\boldsymbol{w},\upsilon), \\
\frac{\partial L_{(2)}}{\partial \upsilon}(\boldsymbol{\xi}) &= \mathbb{E}_x\left[\frac{\partial \ell(\boldsymbol{x}, f_{(1)}(\boldsymbol{x};\boldsymbol{w},\upsilon))}{\partial y}\varphi(\boldsymbol{w}\cdot\boldsymbol{x})\right] = \frac{\partial L_{(1)}}{\partial \upsilon_{(1)}}(\boldsymbol{w},\upsilon).
\end{aligned}$$

The $(\boldsymbol{w}, \upsilon)$-block of $\mathrm{Hess}(\boldsymbol{\xi}_\lambda)$ is given by differentiating the equations above with $(\boldsymbol{w}, \upsilon)$ and thus equal to the Hessian matrix $(\partial^2 L_{(1)}/\partial\boldsymbol{\theta}_{(1)}\partial\boldsymbol{\theta}_{(1)})(\boldsymbol{\theta}^*)$, which is positive definite since $\boldsymbol{\theta}_{(1)} = \boldsymbol{\theta}^*$ is a strict local minimizer of $L_{(1)}(\boldsymbol{\theta}_{(1)})$.

On the other hand, we have

$$\frac{\partial^2 L_{(2)}}{\partial \boldsymbol{u}\partial \boldsymbol{u}}(\boldsymbol{\xi}_\lambda) = \mathbb{E}_x\left[\frac{\partial \ell(\boldsymbol{x},\, f_{(2)}(\boldsymbol{x};\boldsymbol{\xi}_\lambda))}{\partial y}\,\lambda(1-\lambda)\,\upsilon^*\varphi''(\boldsymbol{w}^*\cdot\boldsymbol{x})\,\boldsymbol{x}\boldsymbol{x}^T\right]$$
$$= \lambda(1-\lambda)H,$$

where H is the matrix defined in equation 2.1. Thus, when H is positive definite, the Hessian matrix $\mathrm{Hess}(\boldsymbol{\xi}_\lambda)$ is positive semidefinite if and only if $\lambda \in [0,1]$. This proves that $\boldsymbol{\xi}_\lambda$ is a saddle point of $L_{(2)}(\boldsymbol{\xi})$ for $\lambda \in \mathbb{R}\setminus[0,1]$. Similarly, when H is negative definite, $\boldsymbol{\xi}_\lambda$ is a saddle point of $L_{(2)}(\boldsymbol{\xi})$ for $\lambda \in (0,1)$. When H is indefinite, so is $\mathrm{Hess}(\boldsymbol{\xi}_\lambda)$ for every $\lambda \in \mathbb{R}\setminus\{0,1\}$.

We show that the point $\boldsymbol{\xi} = \boldsymbol{\xi}_\lambda$ is a local minimizer of $L_{(2)}(\boldsymbol{\xi})$ for $\lambda \in (0,1)$ when the matrix H is positive definite. For $\lambda \in (0,1)$, since the matrix $\mathrm{Hess}(\boldsymbol{\xi}_\lambda)$ is positive semidefinite, the Taylor series of $(L_{(2)}(\boldsymbol{\xi}) - L_{(2)}(\boldsymbol{\xi}_\lambda))$ up to the second order is nonnegative in a neighborhood of $\boldsymbol{\xi} = \boldsymbol{\xi}_\lambda$. However, higher-order terms may influence the sign of the Taylor series, since the coefficients of the terms of z^2 and γz are zero for $\gamma = \boldsymbol{w}, \upsilon, \boldsymbol{u}$. Since the terms of $\gamma^k z^\ell$ with $k \geq 2$ are dominated by the term γ^2 near $\boldsymbol{\xi} = \boldsymbol{\xi}_\lambda$, we check if the coefficients of the terms of z^ℓ and γz^ℓ are equal to zero for all $\ell \geq 1$. For all $\boldsymbol{\xi}$ such that $\boldsymbol{u} = \boldsymbol{0}$, $f_{(2)}(\boldsymbol{x};\boldsymbol{\xi}) = \upsilon\varphi(\boldsymbol{w}\cdot\boldsymbol{x})$ is a constant as a function of z, and so is $L_{(2)}(\boldsymbol{\xi})$. Hence, we obtain

$$\frac{\partial^\ell L_{(2)}}{\partial z^\ell}(\boldsymbol{\xi}_\lambda) = 0,$$
$$\frac{\partial^{\ell+1} L_{(2)}}{\partial \gamma\,\partial z^\ell}(\boldsymbol{\xi}_\lambda) = 0, \quad \ell \geq 1,$$

for $\gamma = \boldsymbol{w}, \upsilon$. Therefore, it suffices to check if

$$\frac{\partial^{\ell+1} L_{(2)}}{\partial \boldsymbol{u}\partial z^\ell}(\boldsymbol{\xi}_\lambda) = 0, \quad \ell \geq 1.$$

Since we have already seen that $(\partial L_{(2)}/\partial \boldsymbol{u})(\boldsymbol{\xi}) = 0$ for all $\boldsymbol{\xi}$ such that $\boldsymbol{u} = \boldsymbol{0}$, this equality is confirmed. Finally, we have shown that the coefficients of higher-order terms are all zero, and that they do not influence the sign of the Taylor series. This shows that $(L_{(2)}(\boldsymbol{\xi}) - L_{(2)}(\boldsymbol{\xi}_\lambda))$ is nonnegative near $\boldsymbol{\xi} = \boldsymbol{\xi}_\lambda$, and hence the proof is complete. In the case that the matrix H is negative, it is proved similarly. □

Proof of Theorem 1. Since $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are no longer scalars, we cannot use the coordinate system, given by equation A.1. Therefore, in order to analyze

the Hessian matrix, we introduce another coordinate system $\boldsymbol{\xi} = (\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{u}, \boldsymbol{z})$ as

$$\begin{cases} \boldsymbol{w} = \dfrac{\boldsymbol{w}_1 + \boldsymbol{w}_2}{2} \\ \boldsymbol{v} = \boldsymbol{v}_1 + \boldsymbol{v}_2 \\ \boldsymbol{u} = \dfrac{\boldsymbol{w}_1 - \boldsymbol{w}_2}{2} \\ \boldsymbol{z} = \boldsymbol{v}_1 - \boldsymbol{v}_2 \end{cases}.$$

In this coordinate system, the point $\boldsymbol{\theta}_\lambda$ is denoted as $\boldsymbol{\xi}_\lambda = (\boldsymbol{w}^*, \boldsymbol{v}^*, \boldsymbol{0}, (2\lambda - 1)\boldsymbol{v}^*)$.

Fix $\lambda \in \mathbb{R}$ arbitrarily. We show that the Hessian matrix $\mathrm{Hess}(\boldsymbol{\xi}_\lambda)$ of $L_{(2)}(\boldsymbol{\xi})$ at $\boldsymbol{\xi} = \boldsymbol{\xi}_\lambda$ has both positive and negative eigenvalues. It suffices to show that the $(\boldsymbol{w}, \boldsymbol{v})$-part of the Hessian is positive definite and that the $(\boldsymbol{u}, \boldsymbol{z})$-part is not positive semidefinite. They imply that the full Hessian matrix $\mathrm{Hess}(\boldsymbol{\xi}_\lambda)$ is neither negative semidefinite nor positive semidefinite, and hence that $\mathrm{Hess}(\boldsymbol{\xi}_\lambda)$ is indefinite.

One can check that the $(\boldsymbol{w}, \boldsymbol{v})$-part of $\mathrm{Hess}(\boldsymbol{\xi}_\lambda)$ is equal to the Hessian matrix of $L_{(1)}$ at $\boldsymbol{\theta}^*$ by direct calculation. Since $\boldsymbol{\theta}^*$ is a strict local minimizer of $L_{(1)}$, this is positive definite.

On the other hand, we have

$$\frac{\partial^2 L_{(2)}}{\partial \boldsymbol{z} \partial \boldsymbol{u}}(\boldsymbol{\xi}_\lambda) = \mathbb{E}_{\boldsymbol{x}}\left[\frac{\partial \ell(\boldsymbol{x}, \boldsymbol{f}_{(2)}(\boldsymbol{x}; \boldsymbol{\xi}_\lambda))}{\partial \boldsymbol{y}} \varphi'(\boldsymbol{w}^* \cdot \boldsymbol{x})\, \boldsymbol{x}^T\right], \tag{A.2}$$

$$\frac{\partial^2 L_{(2)}}{\partial \boldsymbol{z} \partial \boldsymbol{z}}(\boldsymbol{\xi}_\lambda) = O, \tag{A.3}$$

and thus the $(\boldsymbol{u}, \boldsymbol{z})$-part B of $\mathrm{Hess}(\boldsymbol{\xi}_\lambda)$ is written as

$$B = \begin{pmatrix} \frac{\partial^2 L_{(2)}}{\partial \boldsymbol{u} \partial \boldsymbol{u}}(\boldsymbol{\xi}_\lambda) & \mathbb{E}_{\boldsymbol{x}}\left[\frac{\partial \ell(\boldsymbol{x}, \boldsymbol{f}_{(2)}(\boldsymbol{x}; \boldsymbol{\xi}_\lambda))}{\partial \boldsymbol{y}} \varphi'(\boldsymbol{w}^* \cdot \boldsymbol{x})\, \boldsymbol{x}^T\right]^T \\ \mathbb{E}_{\boldsymbol{x}}\left[\frac{\partial \ell(\boldsymbol{x}, \boldsymbol{f}_{(2)}(\boldsymbol{x}; \boldsymbol{\xi}_\lambda))}{\partial \boldsymbol{y}} \varphi'(\boldsymbol{w}^* \cdot \boldsymbol{x})\, \boldsymbol{x}^T\right] & O \end{pmatrix},$$

where we treat $\partial \ell / \partial \boldsymbol{y}$ as a column vector. Here, since $\boldsymbol{f}_{(2)}(\boldsymbol{x}; \boldsymbol{\xi}_\lambda) \equiv \boldsymbol{f}_{(1)}(\boldsymbol{x}; \boldsymbol{\theta}^*)$, the nondiagonal block, equation A.2, is equal to the matrix, equation 2.2, and thus is nonzero by assumption. Hence, the $(\boldsymbol{u}, \boldsymbol{z})$-part above is not positive semidifinite. In fact, choosing vectors $\boldsymbol{a} \in \mathbb{R}^m$ and $\boldsymbol{b} \in \mathbb{R}^{n+1}$ such that

$$\boldsymbol{a}^T\, \mathbb{E}_{\boldsymbol{x}}\left[\frac{\partial \ell(\boldsymbol{x}, \boldsymbol{f}_{(2)}(\boldsymbol{x}; \boldsymbol{\xi}_\lambda))}{\partial \boldsymbol{y}} \varphi'(\boldsymbol{w}^* \cdot \boldsymbol{x}) \boldsymbol{x}^T\right] \boldsymbol{b} < 0,$$

the vector $\boldsymbol{c}_\varepsilon := (\varepsilon \boldsymbol{b}^T, \boldsymbol{a}^T)^T$ satisfies $\boldsymbol{c}_\varepsilon^T\, B\, \boldsymbol{c}_\varepsilon < 0$ for sufficiently small $\varepsilon > 0$. Such vectors $(\boldsymbol{a}, \boldsymbol{b})$ always exist—for instance, by letting $a_i = -\rho$, $b_j = 1$,

and other entries be 0, where (i, j) is an index such that (i, j)-entry ρ of the matrix A.2 is nonzero. □

Acknowledgments

I express my gratitude to Akio Fujiwara for his helpful guidance, discussions, and advice. I thank as well Yuzuru Sato for many discussions and helpful comments. And I thank the anonymous referees for their insightful suggestions to improve this article.

References

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

Carr, J. (1981). *Applications of centre manifold theory.* New York: Springer.

Cousseau, F., Ozeki, T., & Amari, S.-I. (2008). Dynamics of learning in multilayer perceptrons near singularities. *IEEE Transactions on Neural Networks*, *19*(8), 1313–1328.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, *2*(4), 303–314.

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

Funahashi, K. (1989). On the approximate realization of continuous mappings by neural networks. *Neural Networks*, *2*(3), 183–192.

Sonoda, S., & Murata, N. (2017). Neural network with unbounded activation functions is universal approximator. *Applied and Computational Harmonic Analysis*, *43*(2), 233–268.

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

Received July 26, 2019; accepted November 26, 2019.