Center manifold analysis of singular regions in learning of three-layer perceptron

大阪大学大学院 理学研究科 数学専攻
筒井大二 (Daiji TSUTSUI)

概要

階層型のニューラルネットワークは，そのパラメータ空間内に，隠れ素子の縮退に由来する特異領域を多く持つ．特に，本稿で扱う三層パーセプトロンは，吸引的な部分と反発的な部分の両方からなる一次元の特異領域を持つ．このような特異領域はしばしば Milnor-like attractor とよばれる．Milnor-like attractor の近傍では，いくつかのパラメータが非常に速く収束し，学習の力学系は残りのパラメータに縮約されることが，経験的に知られていた．我々は，中心多様体理論に基づき，この現象を厳密に解析した．

1 基礎概念

本節では，本稿を通じて用いられる記号を導入するとともに，機械学習に関する基本概念を概観する．

1.1 パーセプトロン

パーセプトロン (perceptron) は，生物学的な脳の構造および機能をモデル化した，人工ニューラルネットワーク (neural network) の一種である．多数の人工ニューロン (artificial neuron) により構成され，それぞれのニューロンは他のニューロンから受け取った信号を重み付きで積算して，それに応じた出力を発する．

単一の人工ニューロン，あるいは単純パーセプトロン (simple perceptron) の入出力関係は，以下のようにモデル化される:

$$y = \varphi\left(\sum_{i=1}^{n} w^i x_i - b\right) = \varphi(\boldsymbol{w}\cdot\boldsymbol{x} - b).$$

ベクトル $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ は入力信号を，$y \in \mathbb{R}$ は出力信号を表す．一方，パラメータ $\boldsymbol{w}, b$ は，人工ニューロンを特徴付けるパラメータである．ベクトル $\boldsymbol{w} = (w^1, \ldots, w^n) \in \mathbb{R}^n$ はニューロンの持つ重みであり，$b \in \mathbb{R}$ はニューロンの発火のしきい値を表すパラメータで，バイアスとよばれる．関数 φ は非線形の関数で，活性化関数 (activation function) とよばれる．図 1 は人工ニューロンの概念図である．

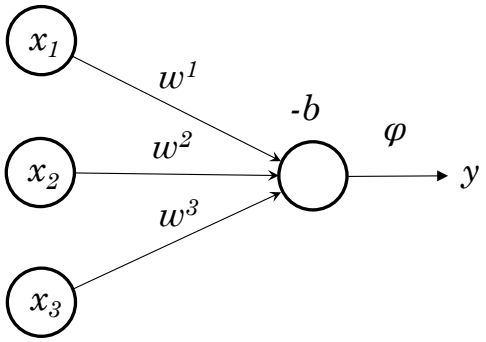

図 1　人工ニューロンの概念図．入力 $\boldsymbol{x}$ を重み $\boldsymbol{w}$ で積算し，バイアス b と比較したのちに，非線形関数 φ を通した値を出力する．

ニューラルネットワーク研究の黎明期においては，McCulloch-Pitts[5] がステップ関数

$$\varphi(z) := \begin{cases} 1 & (z \geq 0) \\ 0 & (z < 0) \end{cases}$$

を活性化関数として採用した．これは，当時の入出力が $\{0,1\}$-値をとるという設定の上では自然な選択であった．しかし，後になって連続値の入出力を用いることが一般的となり，それに伴って活性化関数のクラスも拡張され，今日では多様な関数が活性化関数として用いられている．多くの活性化関数は有限点を除いて微分可能である．これは，後に述べる，機械学習において標準的な学習手法である，勾配降下法を利用するためである．

記号の簡約化のため，ベクトル $\boldsymbol{x}, \boldsymbol{w}$ を以下のように拡大する．

$$\boldsymbol{x} = (x_0, x_1, \ldots, x_n), \qquad \boldsymbol{w} = (w^0, w^1, \ldots, w^n),$$

ここに，$x_0 := 1, w^0 := -b$ であり，この記号のもと，

$$\sum_{i=1}^{n} w^i x_i - b = \boldsymbol{w} \cdot \boldsymbol{x}.$$

と簡潔な表記を得る．本稿では，以下を通してこの記号法を用いる．

一般のパーセプトロンは，多数の層によって構成される階層構造を持つモデルである．ここに，層とは一個以上の人工ニューロンからなる集合を言う．入力信号は，最初の層のニューロンによって受け取られ，それらのニューロンの出力信号が次の層へと送られる．次の層のニューロンはその信号を入力として受け取り，さらに次の層へ出力を送る．我々は，最後の層の出力を受け取り，所与の入力に対する系全体の出力とみなす．このようなパーセプトロンは多層パーセプトロン (multi-layer perceptron) とよばれる．図 2 に，多層パーセプトロンの概念図を示す．

本稿では三層パーセプトロン (three-layer perceptron)，すなわち，入力層，隠れ層，出力層の三つの層からなる多層パーセプトロンを扱う（図 3）．数学的には，以下で与えられる対象である．

定義 1.1 (三層パーセプトロン)**.** n 個の入力素子，d 個の隠れ素子，m 個の出力素子からなる三層

パーセプトロンを $(n\text{-}d\text{-}m)$-パーセプトロンといい，次の入出力関係で定義する:

$$\boldsymbol{f}_{(d)}(\boldsymbol{x};\boldsymbol{\theta}) := \sum_{j=1}^{d} \boldsymbol{v}_j \varphi(\boldsymbol{w}_j \cdot \boldsymbol{x}). \tag{1}$$

ここに，$\boldsymbol{x} = (1, x_1, \ldots, x_n) \in \mathbb{R}^{n+1}$ は入力，$\boldsymbol{\theta} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_d, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_d)$ は系のパラメータであり，$\boldsymbol{w}_1, \ldots, \boldsymbol{w}_d \in \mathbb{R}^{n+1}, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_d \in \mathbb{R}^m$.

以下では「パーセプトロン」と言ったとき，文脈に応じて，パラメータ $\boldsymbol{\theta}$ によって指定される関数 $\boldsymbol{f}_{(d)}(\boldsymbol{x};\boldsymbol{\theta})$ と関数族 $\{\boldsymbol{f}_{(d)}(\boldsymbol{x};\boldsymbol{\theta})\}_{\boldsymbol{\theta}}$ のどちらも指す場合がある．

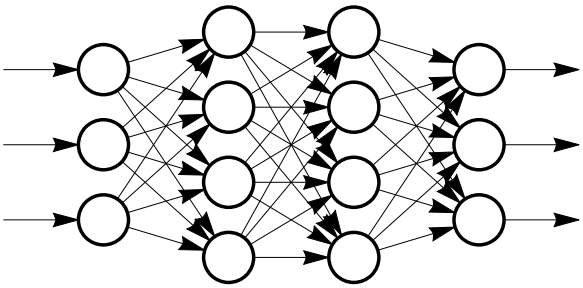

図 2　多層パーセプトロン．

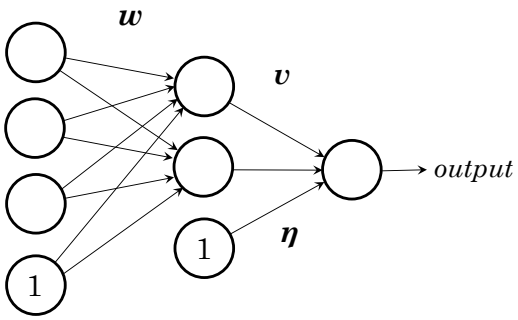

図 3　三層パーセプトロン．

1.2　機械学習

本稿では教師あり学習 (supervised learning) について考察する．学習における各問題設定で，教師関数 (teacher function) とよばれる入出力関係 $T(\boldsymbol{x})$ を用意する．学習の目的は，生徒であるパーセプトロンが所与の教師関数 $T(\boldsymbol{x})$ を可能な限り模倣できるよう，系のパラメータを調節することである．学習の尺度は，例えば平均損失関数 (averaged loss function) $L(\boldsymbol{\theta})$ により表現される．学習とは，平均損失関数を用いて言い換えると，この関数 $L(\boldsymbol{\theta})$ を最小化するパラメータ $\boldsymbol{\theta}$ を探索することである．平均損失関数 $L(\boldsymbol{\theta})$ は次のように定義される．関数 $\ell(\boldsymbol{x}, \boldsymbol{y})$ を，非負値の関数であって，$\boldsymbol{y} = T(\boldsymbol{x})$ のときかつそのときに限り 0 をとるものとする．このような関数を（即時）損失関数 (instantaneous loss function) という．二乗誤差関数

$$\ell(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}||\boldsymbol{y} - T(\boldsymbol{x})||^2$$

は損失関数の典型的な例である．損失関数 $\ell(\boldsymbol{x}, \boldsymbol{y})$ が与えられた上で，平均損失関数 $L(\boldsymbol{\theta})$ は

$$L(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{x}}[\ell(\boldsymbol{x}, \boldsymbol{f}_{(d)}(\boldsymbol{x};\boldsymbol{\theta})]$$

で定義される．ここに，$\mathbb{E}_{\boldsymbol{x}}$ は入力信号 $\boldsymbol{x}$ に関する期待値であり，$\boldsymbol{x}$ は未知の確率分布に従う確率変数であることを仮定している．

本稿では，勾配降下法による学習に焦点を当てる．平均損失関数 $L(\boldsymbol{\theta})$ を最小化するために，我々は微分方程式

$$\dot{\boldsymbol{\theta}} = -\frac{\partial L}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}) \tag{2}$$

を利用することを考える．つまり，関数 $L(\boldsymbol{\theta})$ の勾配の逆方向にパラメータ $\boldsymbol{\theta}$ を動かす力学系を考え，その積分曲線の行き着く先として，$L(\boldsymbol{\theta})$ を最小にする点を探索するのである．このような探索方法は勾配降下法 (gradient descent method) と呼ばれる．計算機で実装するにあたっては，しばしば Euler 法を用いた公式:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \varepsilon \frac{\partial L}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_t)$$

によって $\boldsymbol{\theta}$ を更新する．ここで，ε は学習定数 (learning rate) とよばれる正の数であり，小さな値を想定している．

1.3　プラトー現象

実際の学習過程において，しばしば損失関数がある長い期間減少しなくなり，その後再び減少を開始する，といったことが起きる．そのような現象はプラトー現象 (plateau phenomena) などとよばれる．平均損失関数 $L(\boldsymbol{\theta})$ の臨界点の近くでは，勾配が非常に小さくなるためにプラトー現象が起きることはよく知られている．パラメータ $\boldsymbol{\theta}$ は，長い時間をかけて臨界点から離れていった後，通常の学習を再開するのである．

Fukumizu と Amari[4] は，(n-($d-1$)-1)-パーセプトロンに対応する損失関数 $L_{(d-1)}$ が極小点を持つとき，(n-d-1)-パーセプトロンがその構造に (n-($d-1$)-1)-パーセプトロンを含むことに由来して，損失関数 $L_{(d)}$ のパラメータ空間に特殊な構造が現れることを見出した．これは，しばしば Milnor-like attractor とよばれ，彼らはこの構造により深刻なプラトー現象が引き起こされることを主張した．本稿では，この Milnor-like attractor の近傍で生じる勾配法の力学を解析する．

2　先行研究

パーセプトロンのパラメータ空間は「パーセプトロン多様体」などとよばれることがある．しかしながら通常，パラメータ空間内には同じ入出力関係に対応する点が遍在し，そのためにパラメータ空間は，普通の意味での多様体にはならない．そのような同一の入出力関係に対応する点の集合は，特異領域 (singular region) とよばれる．本節では，Fukumizu と Amari[4] による，三層パーセプトロンの隠れ素子の縮退によって生じる特異領域周辺で起こる学習の停滞現象について概説する．

式 (1) で定義される三層パーセプトロンについて考察する．本稿では，三層パーセプトロンの隠れ素子の縮退を観察するための最も簡単なモデルとして (n-2-m)-パーセプトロン

$$\begin{aligned} \boldsymbol{f}_{(2)}(\boldsymbol{x};\boldsymbol{\theta}) &= \boldsymbol{v}_1\,\varphi(\boldsymbol{w}_1\cdot\boldsymbol{x}) + \boldsymbol{v}_2\,\varphi(\boldsymbol{w}_2\cdot\boldsymbol{x}), \\ \boldsymbol{\theta} &= (\boldsymbol{w}_1,\boldsymbol{w}_2,\boldsymbol{v}_1,\boldsymbol{v}_2). \end{aligned} \tag{3}$$

を考察する．

重みパラメータ $\boldsymbol{w}_1,\boldsymbol{w}_2$ が $\boldsymbol{w}_1=\boldsymbol{w}_2=\boldsymbol{w}$ を満たすとき，このモデルは

$$\boldsymbol{f}_{(2)}(\boldsymbol{x};\boldsymbol{\theta}) = (\boldsymbol{v}_1+\boldsymbol{v}_2)\,\varphi(\boldsymbol{w}\cdot\boldsymbol{x}) = \boldsymbol{f}_{(1)}(\boldsymbol{x};(\boldsymbol{w},\boldsymbol{v})).$$

のように退化し，パラメータ $(\boldsymbol{w}, \boldsymbol{v})$ によって特徴付けられる (n-1-m)-パーセプトロンと同じく振る舞う．ここに，$\boldsymbol{v} := \boldsymbol{v}_1 + \boldsymbol{v}_2$ とした．換言すると，$\boldsymbol{w} \in \mathbb{R}^{n+1}, \boldsymbol{v} \in \mathbb{R}^m$ に対し，(n-2-m)-パーセプトロンのパラメータ空間における部分集合

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

は特異領域である．

隠れ素子の個数 d に着目し，(n-d-m)-パーセプトロン $\boldsymbol{f}_{(d)}(\boldsymbol{x}; \boldsymbol{\theta})$ に対する損失関数を $L_{(d)}(\boldsymbol{\theta})$ で表す．すなわち，

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

本稿で主に扱われるのは，$m = 1$，すなわち出力信号が 1 次元の場合である．このとき，次の結果が知られている．

命題 2.1 (Fukumizu and Amari[4])**.** パラメータ $\boldsymbol{\theta}^* = (\boldsymbol{w}^*, v^*)$ を，$L_{(1)}$ の狭義極小点とする．各 $\lambda \in \mathbb{R}$ に対し，

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

とおく．このとき各 $\lambda \in \mathbb{R}$ で，$\boldsymbol{\theta}_\lambda$ は関数 $L_{(2)}$ の臨界点である．さらに，H を

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

で定まる $(n+1) \times (n+1)$ 行列とすると，H が正定値（負定値）ならば，点 $\boldsymbol{\theta}_\lambda$ は $\lambda \in (0, 1)$ に対して $L_{(2)}$ の極小点（鞍点）であり，$\lambda \in \mathbb{R} \setminus [0, 1]$ に対して鞍点（極小点）である．また，H が不定値ならば，各 $\lambda \in \mathbb{R}$ に対して点 $\boldsymbol{\theta}_\lambda$ は $L_{(2)}$ の鞍点である．

この命題は，(n-2-1)-パーセプトロンの勾配法による学習において，次のような現象が起き得ることを示唆している．パラメータ空間内の一次元領域 $R(\boldsymbol{w}^*, v^*) = \{\, \boldsymbol{\theta}_\lambda \mid \lambda \in \mathbb{R} \,\}$ は，上の命題が成立する条件のもと，$L_{(2)}$ の極小点からなる吸引的な部分と，鞍点からなる反発的な部分の両方を持つ．この一次元領域は Milnor-like attractor などとよばれる．パラメータ $\boldsymbol{\theta}$ は，学習の過程で吸引的な部分の近くを通るときにこの領域に捕捉される．力学系 (2) を考える上では，平衡点に収束して $\boldsymbol{\theta}$ は止まってしまうが，実際の学習においては，（例えば期待値を推定値で代用するために）パラメータ $\boldsymbol{\theta}$ の挙動に確率的な揺動が生じる．長い時間この一次元領域の中を揺れ動いて反発的な部分に達した後，ようやくこの領域から脱出して学習が再開する．

ここに述べた構造は，隠れ素子の個数 d については普遍的である．Fukumizu と Amari[4] では，(n-d-1)-パーセプトロンが (n-($d-1$)-1)-パーセプトロンに縮退する設定のもとで解析を行っており，上で述べた命題は $d = 2$ の場合に相当する．

$m \geq 2$ の場合についての結果も述べておく．この場合も $m = 1$ の場合と同様に，$L_{(2)}$ の臨界点からなる一次元の特異領域が存在する．しかしこの場合には，特異領域の各点は反発的であり，吸引的な部分は存在しない．ゆえに，上のような Milnor-like attractor の構造はこの場合には現れない．

命題 2.2. パラメータ $\boldsymbol{\theta}^* = (\boldsymbol{w}^*, \boldsymbol{v}^*)$ を $L_{(1)}$ の極小点とする．各 $\lambda \in \mathbb{R}$ に対し，

$$\boldsymbol{\theta}_\lambda := (\boldsymbol{w}^*, \boldsymbol{w}^*, \lambda \boldsymbol{v}^*, (1-\lambda) \boldsymbol{v}^*) \tag{4}$$

とおく．$m \times (n+1)$ 行列

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

が非ゼロ行列ならば，各 $\lambda \in \mathbb{R}$ に対し，点 $\boldsymbol{\theta}_\lambda$ は $L_{(2)}$ の鞍点である．ただしここで，$\partial\ell/\partial\boldsymbol{\theta}$ を列ベクトルとみなしている．

3　主結果

Wei ら [6] は，(n-2-1)-パーセプトロンの解析の中で，次のようなパラメータ空間の座標変換を導入した:

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

Amari ら [1] は，Milnor-like attractor の周辺では，パラメータ $(\boldsymbol{w}, v)$ は平衡点 $(\boldsymbol{w}^*, v^*)$ に素早く収束すると主張した．ここに，$(\boldsymbol{w}^*, v^*)$ は命題 2.1 内で述べた $L_{(1)}$ の極小点である．彼らは，その仮説のもと，学習の力学系をパラメータ $(\boldsymbol{u}, z)$ のものへと縮約して解析を行った．

本稿の主結果は，新たな座標変換を導入することで，Amari らの仮説を中心多様体の意味で正当化したことである．その結果として，中心多様体の観点から厳密な縮約力学系を得ることに成功した．主結果は以下のように述べられる．

3.1　特異領域の中心多様体

定理 3.1. (n-2-1)-パーセプトロン $f_{(2)}(\boldsymbol{x};\boldsymbol{\theta})$ のパラメータ空間上の新たな座標系 $\boldsymbol{\xi} = (\boldsymbol{w}, v, \boldsymbol{u}, z)$ を以下で定める．

$$\begin{cases} \boldsymbol{w} = \dfrac{v_1\left(\boldsymbol{w}_1 - \boldsymbol{w}^*\right) + v_2\left(\boldsymbol{w}_2 - \boldsymbol{w}^*\right)}{v^*} + \boldsymbol{w}^* \\ v = v_1 + v_2 \\ \boldsymbol{u} = \dfrac{v_2\left(\boldsymbol{w}_1 - \boldsymbol{w}^*\right) - v_1\left(\boldsymbol{w}_2 - \boldsymbol{w}^*\right)}{v^*} \\ z = v_1 - v_2 \end{cases}. \tag{5}$$

ただしここに，$\boldsymbol{\theta}^* = (\boldsymbol{w}^*, v^*)$ は $L_{(1)}$ の狭義極小点．この座標系 $\boldsymbol{\xi}$ のもと，力学系 (2) は平衡点 $\boldsymbol{\theta} = \boldsymbol{\theta}_0, \boldsymbol{\theta}_1$ の周りで中心多様体構造を許容し，パラメータ $(\boldsymbol{w}, v)$ は指数的に速く平衡点 $(\boldsymbol{w}^*, v^*)$ に収束する．

中心多様体構造が許容される二点 $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1$ は，Milnor-like attractor $\{\,\boldsymbol{\theta}_\lambda \mid \lambda \in \mathbb{R}\,\}$ の，吸引的な部分と反発的な部分の境界にあたる点であり，Milnor-like attractor 構造によって引き起こされる力学

のモードが変化する点である．上の定理は，その点の周りで中心多様体縮約された力学系を解析できることを保証している．

定理 3.1 により，パラメータ $\boldsymbol{\theta}$ が $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1$ に近いという仮定のもと，力学系 (2) は，中心多様体理論 [2] における標準的な方法で縮約される．これらの二点 $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1$ は (5) で定まる座標系 $\boldsymbol{\xi} = (\boldsymbol{w}, v, \boldsymbol{u}, z)$ で表すと

$$\boldsymbol{\xi}_0 = (\boldsymbol{w}^*, v^*, \boldsymbol{0}, -v^*), \qquad \boldsymbol{\xi}_1 = (\boldsymbol{w}^*, v^*, \boldsymbol{0}, v^*)$$

であるが，それらの周りでは，次のような縮約力学系を得る．点 $\boldsymbol{\xi} = \boldsymbol{\xi}_0$ の周りでは，

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

これは高次の項を無視すると陽に積分でき，

$$||\boldsymbol{u}||^2 = (z + v^*)^2 + C, \tag{6}$$

なる軌跡に従うことがわかる．ここに，C は積分定数．

同様に，点 $\boldsymbol{\xi} = \boldsymbol{\xi}_1$ の周りでは，

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

と縮約され，これを積分して

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

なる軌跡を得る．

3.2 証明の概略

中心多様体の存在定理（例えば [2] を参照）により，(5) で定まる座標系 $\boldsymbol{\xi}$ のもとで，力学系 (2) を点 $\boldsymbol{\xi} = \boldsymbol{\xi}_0, \boldsymbol{\xi}_1$ の周りで線形化した係数行列（Jacobi 行列）が

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

のような形に分解し，行列 A の全ての固有値の実部が負，B の全ての固有値の実部がゼロであることを示せば十分である．

直接の計算により，$\boldsymbol{\xi} = \boldsymbol{\xi}_0, \boldsymbol{\xi}_1$ において係数行列が上のように分解して，B がゼロ行列になることがわかる．また，力学系 (2) は座標系 $\boldsymbol{\xi}$ のもと

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

と変換され，$\boldsymbol{\xi} = \boldsymbol{\xi}_0$ が $L_{(2)}$ の臨界点であることから，係数行列は

$$\begin{aligned}\frac{\partial \dot{\boldsymbol{\xi}}}{\partial \boldsymbol{\xi}}(\boldsymbol{\xi}_0) &= -\frac{\partial}{\partial \boldsymbol{\xi}} \left\{ \frac{\partial \boldsymbol{\xi}}{\partial \boldsymbol{\theta}} \left(\frac{\partial \boldsymbol{\xi}}{\partial \boldsymbol{\theta}} \right)^T \left(\frac{\partial L_{(2)}}{\partial \boldsymbol{\xi}}(\boldsymbol{\xi}) \right) \right\} \Bigg|_{\boldsymbol{\xi}=\boldsymbol{\xi}_0} \\ &= -\left(\frac{\partial \boldsymbol{\xi}}{\partial \boldsymbol{\theta}} \left(\frac{\partial \boldsymbol{\xi}}{\partial \boldsymbol{\theta}} \right)^T \right) \frac{\partial^2 L_{(2)}}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}}(\boldsymbol{\xi}_0)\end{aligned}$$

の形である．命題 2.1 を示す際の計算により，$\boldsymbol{\xi} = \boldsymbol{\xi}_0$ で Hesse 行列 $\partial^2 L_{(2)}/\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}$ が半正定値であることがわかり，さらに $(\partial \boldsymbol{\xi}/\partial \boldsymbol{\theta})(\partial \boldsymbol{\xi}/\partial \boldsymbol{\theta})^T$ が正定値であることから，係数行列 $\partial \dot{\boldsymbol{\xi}}/\partial \boldsymbol{\xi}$ の固有値は全て非正の実数であることが従う．他方，上の関係から，$\boldsymbol{\xi} = \boldsymbol{\xi}_0$ における係数行列 $\partial \dot{\boldsymbol{\xi}}/\partial \boldsymbol{\xi}$ の階数は Hesse 行列の階数と一致する．ゆえに，Hesse 行列の階数を調べることで，部分行列 A の固有値はゼロを含まず，全て負の実数であることが示される．

参考文献

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

[3] F. Cousseau, T. Ozeki, and S.-I. Amari, “Dynamics of learning in multilayer perceptrons near singularities,” IEEE Transactions on Neural Networks, **19**(8), 1313-1328 (2008).

[4] K. Fukumizu and S.-I. Amari, “Local minima and plateaus in hierarchical structures of multilayer perceptrons,” Neural Networks, **13**(3), 317-327 (1999).

[5] W. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” Bulletin of Mathematical Biophysics, **5**, 115-133 (1943).

[6] 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).