甲南大学　博士学位論文

数学の学習における過剰般化現象のシミュレーションに関する研究

甲南大学大学院

自然科学研究科　知能情報学専攻

2024年3月

学籍番号 31523001

鷲野朋広

概要

情報科学における学習理論では，モデル学習の初期パラメータを連続的に変化させたとき，理解の状態変化を表すダイナミクスがいくつかの種類に分類されることが知られている．この研究成果を数学の学習における学習者の理解の状態（理解の様相）を可視化する方法として応用することを試みた．

数学の学習において A と B の 2 つの概念が共通概念を含む場合，関連付けて理解することによって全体を理解できるものがある．また，一方だけではなく関連付けて理解することで深い理解に結びつくことがある．この A と B に該当するものの一例として高等学校数学科の単元「場合の数と確率」における「順列」と「組合せ」がある．学習者が「順列」と「組合せ」の 2 つの概念を理解する過程では，「過剰般化」（特定の規則や意味的特徴を過剰に一般化してしまう現象）が生じることがある．この「過剰般化」は，「順列」と「組合せ」の 2 つの概念を理解する過程において生徒が誤った理解をする一因と考えられる．

本研究では，「順列と組合せ」の学習の際に生じる過剰般化の現象を学習者のテストの点数をもとに，情報科学におけるニューラルネットワークの手法を用いて，損失曲面上に学習者の理解の状態を可視化する方法を考案し，数学の学習における過剰般化現象のシミュレーションを行う方法を確立した．

A study on simulations of the phenomenon of overgeneralization in learning mathematics

Abstract

In learning theory in information science, it is known that when the initial parameters of model learning are continuously changed, the dynamics that represent changes in the state of understanding can be classified into several types. We attempted to apply the results of learning theory as a method for visualizing the state of learners' understanding (aspect of understanding) in learning mathematics.

When two concepts A and B contain a common concept, it is possible to understand all the concepts by understanding them in relation to each other. In addition, understanding concepts A and B not just individually but also in relation to each other can lead to a deeper understanding. "Permutations" and "combinations" in the unit "The Number of Cases and Probability" in high school mathematics is an example of such A and B. In the process of learners gaining an understanding of the two concepts of permutations and combinations, "overgeneralization" (the phenomenon of overgeneralizing specific rules or semantic features) may occur. This "overgeneralization" may contribute to students gaining an incorrectly understanding of the two concepts "permutation" and "combination" in the process of understanding them.

In this study, we devised a method to visualize the phenomenon of overgeneralization that occurs when learners study "permutations and combinations" based on learners' test scores.

目次

第 1 章

本研究の背景と目的

英語学習において白畑・若林・村野 ([1]) によれば，日本語を母語とする英語学習者が第二言語習得をするために第二言語のある規則をその規則が適用できない他の項目まで当てはめてしまったために生じた誤りの原因を「過剰般化 (overgeneralization)」としている．

数学の学習において 2 つの概念 A，B が内容的に似ている概念である場合，A，B をこの順に学習したとき，先に学習した A により後に学んだ B の概念理解への影響が考えられる．これを「B を A とする過剰般化」と呼ぶ．また逆に，B の学習後に少し時間が経過した段階で強化学習として A を再学習する際に生じる「A を B とする過剰般化」 も考えることができる．高等学校数学科「数学 A」の「場合の数と確率」の単元では，一般的に「順列」から「組合せ」への順序で学習を行っている．順列の理解における「選んで並べる」の「選んで」という概念は，後に学ぶ組合せの概念を学習した後で順列の理解に影響を与えることがある．

一般に，以前の学習が新しい学習に影響を及ぼすことを，以前の学習が新しい学習に転移したという．新しい学習を促進する方向で影響を与える時は「正の転移」，新しい学習を阻害する方向で働く時は「負の転移」と呼ばれる．順列から組合せへと学習することにより負の転移現象を考えたとき，順列の学習後組合せの理解が下がってしまう状態が想定される．また組合せの学習後に順列の再学習をすることにより，組合せの理解に影響を与えて順列の理解が下がってしまう状態も想定される．このような数学の学習における過剰般化の現象を数学教育として数理科学的に明らかにしたい．

情報科学における学習理論を数学的に解明した研究として，代数幾何に関連する数学的な概念を学習システムの数理として解明している文献がある ([2])．この理論および手法を数学教育の研究として発展させるためには実践研究を可能にする方法を開発しなければならない．本研究で用いる情報科学における学習理論研究の状況は以下の状況である．

階層構造を持ち非線形なニューラルネットワークは，正則モデルとは異なり特異モデルと呼ばれている．特異モデルでは，真の分布を実現する最適なパラメータが 1 点には定まらず，次元をもつ広がりとなり解析的集合をなす．真の分布より学習モデルが少ないユニット数で実現できる場合，真の分布を実現するパラメータが多様体ではなく特異点を含んでいる．特定のデータにモデルが過剰に適合（学習）してしまうことを過学習という．たとえば，２つの教科のテストの結果をそれぞれ縦軸と横軸とし，実際の点数を記入して散布図を作成し，予測パターンを曲線として書き込む場合，適正な予測モデルでは全体の分布をバランスよく捉えた緩やかな曲線になる．しかし，過学習が起きると過度にデータにフィットしすぎ，細かく湾曲する線となる．この場合，前者の緩やかな曲線が求めたい適正な予測モデルであり，後者は過学習によって汎用性が失われている状態という（定義 28 参照）．特異モデルに臨界点 (鞍点, 極小点, 極大点) が生じる現象は，学習が停滞し過学習などが起きる原因となっている ([3]).

渡辺 ([4]) は真のパラメータの解析的集合には特異点が含まれていることを示し，特異点の複雑さについて真の分布の中間ユニット数 H_0 が小さい場合特異点はより複雑になり，学習サンプルの数が増加するにつれて特異点はより単純になることを示した．また，特異モデルにおけるベイズ推定に代数幾何の方法を用いて成果をあげ，代数的に特異点を定義して特異点解消定理を用いて学習理論に代数幾何学を応用した．さらに，統計的なベイズ理論を用いてカルバック情報量や汎化誤差や確率的複雑さを数学的に研究した ([3]).

渡辺，福水，萩原，甘利 ([5]) は，多層パーセプトロンではモデル選択の基準である AIC(赤池情報量基準) がよい性質を与えないことを発見した．渡辺 ([6]) はベイズ統計を用いて特異モデルにおいても使える情報量規定について研究した．福水 ([7]) は，パーセプトロンの特異点における対数尤度の解析を行った．

中間ユニット数の一般化については，甘利，福水 ([8]) が中間ユニットが $H = n$ の学習モデルが中間ユニット数 $H = n - 1$ の真の分布を実現する場合を研究している．福水と甘利は $H = n$ の中間ユニットのパラメータ空間に $H_0 = n - 1$ の中間ユニットが埋め込まれた 3 層ニューラルネットワークのパラメータ空間について学習が停滞するプラトー現象が起きることを明らかにした．さらに，小さなニューラルネットワークの大域的な最小値に対応する臨界点の部分集合が，大きなニューラルネットワークの局所的な最小値または鞍点になる可能性があり，プラトー現象の主なメカニズムであることを明らかにした．

ニューラルネットワークのような階層構造をもつモデルでは真のパラメータの集合は複雑な特異点を含むため，その挙動を解析して理論的に議論することが困難である．学習モデルを基に逐次的にパラメータを変えて損失関数を最小にする過程を学習と呼ぶ．特異点はニューラルネットワークの学習ダイナミクスに影響を与えプラトー現象を引き起こす

ことがある．そのため，ニューラルネットワークにおいて深刻な問題となっている．また，甘利 ([9]) は，特異モデルのパラメータ空間はリーマン空間であること，特異点上ではリーマン計量が縮退すること等を示し，自然勾配学習法を用いて特異モデルにおける統計的推定において大きな成果をあげた．

中間ユニット数が 2 である学習モデルに対して，真の分布が学習モデルによって実現される場合と実現されない場合がある．学習モデルの中間ユニット数が 2 から 1 に変化する領域 (Overlap singularity, Elimination singularity) は特異領域と呼ばれている．甘利 ([10]) は Overlap singularity 現象と Elimination singularity 現象の近くの学習ダイナミクスについて研究した．真の分布が特異点上にある場合や特異点に近い場合や正則な点にある場合を調べた．標準正規分布 $\psi(x)$ に対して 2 成分の混合ガウス分布 $c\psi(x-\mu_1)+(1-c)\psi(x-\mu_2)$ におけるパラメータは (c, μ_1, μ_2) である．ここで μ_1, μ_2 の代わりに 2 つの山の重心 (分布の平均値) を示す $v = c\mu_1 + (1-c)\mu_2$ と，山の位置 (平均値の大きさ) の差を示す $a = \mu_2 - \mu_1$ を新しい変数として導入している．甘利 ([11]) は学習モデルの座標系 $\theta = (w_1, w_2, w_3, w_4)$ を新しい座標系 $\xi = (a, b, v, w)$ に変換した．学習する過程においてパラメータ (v, w) の変化が速く，パラメータ (a, b) の変化が遅いことが知られている．そこで (v, w) を最適解 (v^*, w^*) として固定してパラメータ (a, b) の軌道を研究した．甘利らによって特異領域についての安定性や学習の軌道が式で求められている ([12]，[13]，[14])．

Guo らは Overlap singularity 現象と Elimination singularity 現象の近くの学習のダイナミクスを 5 つのパターン (Overlap singularity, Elimination singularity, Cross elimination singularity, Near elimination singularity, Output weight 0) に分類した ([15])．

渡辺や甘利の理論 ([16]) を数学教育の現象に対し具体的に応用させた研究はまだ行われていない．本研究では，最も基本的な例として中間ユニット数が 2 から 1 に変化する特異領域において，3 層ニューラルネットワークにおいて重み (パラメータ) に教育活動としての意味付けを行う．甘利によると，そのような構造は一般の深層回路にその一部分として到る所に埋め込まれていることが知られている．

本研究では初めに，中間ユニットが $H = n$ の学習モデルが中間ユニット数 $H = m$ の真の分布を実現する場合について考察し，代数的集合のパラメータ表示と次元を求め，カルバック情報量や汎化誤差，学習係数を計算する方法を示す．

数学の学習において A と B の 2 つの項目の概念が共通概念を含む場合，一方だけではなく，それに関連付けて理解することで深い理解に結びつくことがある．中原 ([17]) は概念を構成する過程で共通概念経路について研究し，最適な学習の順路を調べた．高等学校

数学科の「場合の数と確率」の単元において，「順列」と「組合せ」の 2 つの概念を理解する過程において生徒が誤る一因でもある「過剰般化」について取り上げ，その現象に対して，ニューラルネットワークを用いた損失曲面により数理モデリングを行い，過剰般化基づいた特異領域である Overlap singularity 現象と Elimination singularity 現象について，シミュレーションを行うことによって，数学教育における活動を可視化することを目指した．学習を行う初期値を学習前の生徒の学習状況として捉え，真の分布を学習後の生徒の学習状況と設定し，学習のダイナミクスを考察する．このとき初期値をずらすことで，シミュレーションすることが可能になる (図 1.1).

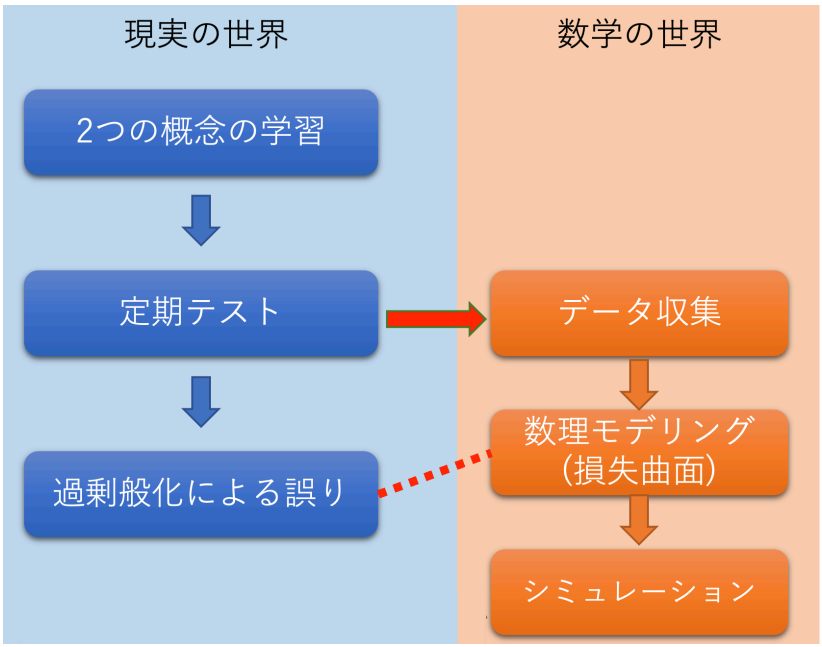


図 1.1　本研究の概要

本研究では，「順列と組合せ」の学習の際に生じる過剰般化の現象を学習者のテストの点数をもとに，情報科学におけるニューラルネットワークの手法を用いて，損失曲面上に学習者の理解の状態を可視化する方法を考案し，数学の学習における過剰般化現象のシミュレーションを行う方法を確立する．

第 2 章

過剰般化現象

ニューラルネットワークの定義をする．その上で中間ユニット数 $H = n \to H_0 = n-1$ の場合，臨界点の埋め込み写像について示す．さらに，過剰般化現象についての考察を行い，順列と組合せ分野への応用の方法を示し，分類測度による可視化について述べる．学習・汎化損失については第 6 章に，数学教育における応用については今後の課題として第 10 章に後述する．

2.1 $M_{\hat{k}}$-ニューラルネットワーク

情報科学において，入力データから真のデータを推論する学習モデルとしてニューラルネットワークを用いる．情報科学における学習を入力データから真のデータに近づけ，ニューラルネットワークの出力パラメータを更新することと定める．

N+2 層のニューラルネットワークを考える．L 個の入力ユニット，1 個の出力ユニット，第 k 層には M_k 個のユニットがあると仮定する ([18]).

本研究では以下の定義 1 から定義 7 に示す手法を用いて定式化する（[19]).

定義 1 $1 \leq k \leq N$ に対して, 第 k 層の j_k 番目の入力ユニット $X_{j_k}^{(k)}$ を次で定める.

$$X_{j_k}^{(k)} := \sum_{j_{k-1}}^{M_{k-1}} w_{j_k j_{k-1}}^{(k)} y_{j_{k-1}}^{(k-1)} + \nu_{j_k}^{(k)}.$$

ここで, $w_{j_k j_{k-1}}^{(k)}$ は第 $k-1$ 層の j_{k-1} 番目ユニットから第 k 層の j_k 番目ユニットへの重み, $y_{j_{k-1}}^{(k-1)}$ は第 $k-1$ 層の j_{k-1} 番目ユニットからの出力, $\nu_{j_k}^{(k)}$ は第 k 層の j_k 番目ユニットの閾値と定める.

また, ベクトルを用いて表す.

$$X_{j_k}^{(k)} = \mathbf{w}_{j_k}^{(k)^T}\mathbf{y}^{(k-1)} + \nu_{j_k}^{(k)} = \tilde{\mathbf{w}}_{j_k}^{(k)^T}\tilde{\mathbf{y}}^{(k-1)}$$

ここで,
$\mathbf{w}_{j_k}^{(k)} = (w_{j_k1}^{(k)}, \ldots, w_{j_kM_{k-1}}^{(k)})^T$, $\tilde{\mathbf{w}}_{j_k}^{(k)^T} = (\mathbf{w}_{j_k}^{(k)}, \nu_{j_k}^{(k)})^T$, $\mathbf{y}^{(k)} = (y_1^{(k)}, \ldots, y_{M_k}^{(k)})$,
$\tilde{\mathbf{y}}^{(k)} = (\mathbf{y}^{(k)^T}, 1)^T$, $\tilde{\mathbf{y}}^{(0)} = (\mathbf{y}^{(0)^T}, 1)^T$, $\mathbf{y}^{(0)} = \mathbf{x}$, $\tilde{\mathbf{x}} = (\mathbf{x}^T, 1)^T$, $\mathbf{x} \in \mathbb{R}^{\mathbf{L}}$.

定義 2 $1 \leq j_k \leq M_k$, $1 \leq k \leq N$ に対して, 第 k 層の j_k 番目の出力ユニット $y_{j_k}^{(k)}$ を次で定める.

$$y_{j_k}^{(k)} := \varphi(X_{j_k}^{(k)}).$$

ここで, $\varphi(x) = \tanh x$ とし, 活性化関数と呼ぶ.

定義 3 入力ベクトル $\mathbf{x} \in \mathbb{R}^{\mathbf{L}}$, パラメータ θ に対して, ニューラルネットワークの出力値 $f(\mathbf{x};\theta)$ を次で定める.

$$f(\mathbf{x};\theta) := \sum_{j_N}^{M_N} v_{j_N}^{(N)} y_{j_N}^{(N)} + \nu^{(N+1)} = \mathbf{v}^{(N)^T}\mathbf{y}^{(N)} + \nu^{(N+1)} = \tilde{\mathbf{v}}^{(N)^T}\tilde{\mathbf{y}}^{(N)}.$$

ここで, $v_{j_N}^{(N)} \in \mathbb{R}$ は, 第 N 層の j_N 番目ユニットから第 $N+1$ 層の出力ユニットへの重み, $\nu^{(N+1)} \in \mathbb{R}$ は出力ユニットの閾値と定める.
また,

$$\mathbf{v}^{(N)} = (v_1^{(N)}, \ldots, v_{M_N}^{(N)})^T, \tilde{\mathbf{v}}^{(N)} = (\mathbf{v}^{(N)^T}, \nu^{(N+1)})^T.$$

θ はニューラルネットワークのパラメータ (重み, 閾値) を全て並べたベクトルと定める.

図 2.1 は $(N+2)$ 層ニューラルネットワークを表している.

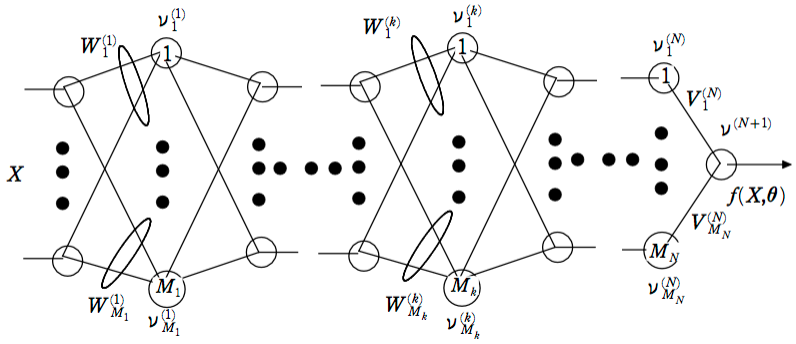


図 2.1　$(N+2)$ 層ニューラルネットワーク

定義 4　k 個の訓練データ $\{(\mathbf{x}^{(m)}, y^{(m)} \in \mathbb{R}^L \times \mathbb{R} |\ m = 1, \ldots, K)\}$ に対して, 誤差関数を次で定める.

$$E(\theta) := \sum_{m=1}^{K} l(y^{(m)}, f(\mathbf{x^{(m)}}; \theta)) \in \mathbb{R}$$

ここで, $l(y^{(m)}, z) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. を損失関数と呼ぶ.

ここで, 学習比率 δ, $i = 1, 2, \ldots$ に対して, 次の等式を用いてパラメータを更新して, 誤差関数を最小にするパラメータを求める.

$$\theta_{i+1} = \theta_i - \delta \frac{\partial E(\theta_i)}{\partial \theta}.$$

θ_* が大域的な最小値のとき, $\frac{\partial E_(\theta_*)}{\partial \theta} = 0$ が成り立ち, 学習が止まる. しかし, 極小値, 極大値, 鞍点のような臨界点で学習が停滞する.

定義 5　第 $\hat{k}$ 層に $M_{\hat{k}}$ 個のユニットを持つニューラルネットワークを $M_{\hat{k}}$-ニューラルネットワークと呼ぶ.

定義 6　パラメータ

$$\theta^{(M_{\hat{k}})} = (\theta_1^{(M_{\hat{k}})}, \ldots, \theta_R^{(M_{\hat{k}})}) \in \Theta_{M_{\hat{k}}}$$

が次の条件を満たすとき，誤差関数 $E_{M_{\hat{k}}}(\theta^{(M_{\hat{k}})})$ の臨界点であるという．

$$\frac{\partial E_{M_{\hat{k}}}(\theta^{(M_{\hat{k}})})}{\partial \theta^{(M_{\hat{k}})}} = \left(\frac{\partial E_{M_{\hat{k}}}(\theta^{(M_{\hat{k}})})}{\partial \theta_1^{(M_{\hat{k}})}}, \ldots, \frac{\partial E_{M_{\hat{k}}}(\theta^{(M_{\hat{k}})})}{\partial \theta_R^{(M_{\hat{k}})}} \right)^T = 0.$$

ここで R は $M_{\hat{k}}$-ニューラルネットワークのパラメータの数, $\Theta_{M_{\hat{k}}}$ はパラメータ全体の集合とする．

定義 7　第 $\hat{k}$ 層に 2 番目から $M_{\hat{k}}$ 番目までの $M_{\hat{k}}-1$ 個のユニットを持つニューラルネットワークを $(M_{\hat{k}}-1)$-ニューラルネットワークと呼ぶ.

このとき, 出力値は

$$f^{(M_{\hat{k}}-1)}(\mathbf{x}; \theta^{(\mathbf{M_{\hat{k}}-1})}) = \sum_{\mathbf{j_N}}^{\mathbf{M_N}} \mathbf{v}_{\mathbf{j_N}}^{(\mathbf{N})} \mathbf{y}_{\mathbf{j_N}}^{(\mathbf{N})} + \nu^{(\mathbf{N+1})}.$$

出力ユニットは

$$y_{j_N}^{(N)} = \varphi(X_{j_N}^{(N)}).$$

入力ユニットは

$$X_{j_{k+1}}^{(k+1)} = \begin{cases} \displaystyle\sum_{j_k=1}^{M_k} w_{j_{k+1}j_k}^{(k+1)} y_{j_k}^{(k)} + \nu_{j_{k+1}}^{(k+1)} (1 \le j_{k+1} \le M_{k+1}, 1 \le k \le N-1, k \ne \hat{k}) \\ \displaystyle\sum_{j_{\hat{k}}=2}^{M_{\hat{k}}} s_{j_{\hat{k}+1}j_{\hat{k}}}^{(\hat{k})} \varphi(\tilde{\mathbf{p}}_{j_{\hat{k}}}^{(\hat{k})^T} \tilde{\mathbf{y}}^{(\hat{k}-1)}) + \xi_{j_{\hat{k}+1}}^{(\hat{k}+1)}, \ (k = \hat{k}) \end{cases}$$

と表される．

ここで, $p_{j_{\hat{k}}j_{\hat{k}-1}}^{(\hat{k})}$ は, 第 $\hat{k}-1$ 層の $j_{\hat{k}-1}$ 番目ユニットから第 $\hat{k}$ 層の j_k 番目ユニットへの重み，$s_{j_{\hat{k}+1}j_{\hat{k}}}^{(\hat{k})}$ は, 第 $\hat{k}+1$ 層の $j_{\hat{k}+1}$ 番目ユニットから第 $\hat{k}$ 層の $j_{\hat{k}}$ 番目ユニットへの重み，$\tilde{\mathbf{p}}_{j_{\hat{k}}}^{(\hat{k})} = (\mathbf{p}_{j_{\hat{k}}}^{(\hat{k})^T}, \tau_{j_{\hat{k}}}^{(\hat{k})})^T$ について, $\mathbf{p}_{j_{\hat{k}}}^{(\hat{k})} = (p_{j_{\hat{k}}1}^{(\hat{k})}, \ldots, p_{j_{\hat{k}}M_{\hat{k}-1}}^{(\hat{k})})^T$ は $j_{\hat{k}}$ 番目ユニットの第 $\hat{k}-1$ 層と第 $\hat{k}$ 層の間の重み, $\tilde{\mathbf{s}}_{j_{\hat{k}}}^{(\hat{k})} = (\mathbf{s}_{j_{\hat{k}}}^{(\hat{k})^T}, \xi_{j_{\hat{k}+1}}^{(\hat{k}+1)})$ について, $\mathbf{s}_{j_{\hat{k}}}^{(\hat{k})} = (s_{1j_{\hat{k}}}^{(\hat{k})}, \ldots, s_{M_{\hat{k}+1}j_{\hat{k}}}^{(\hat{k})})^T$ は $j_{\hat{k}}$ 番目ユニットの第 $\hat{k}$ 層と第 $\hat{k}+1$ 層の間の重み，$\tau_{j_{\hat{k}}}^{(\hat{k})}$ は第 $\hat{k}$ 層の $j_{\hat{k}}$ 番目ユニットの閾値, $\xi_{j_{\hat{k}+1}}^{(\hat{k}+1)}$ は第 $\hat{k}+1$ 層の $j_{\hat{k}+1}$ 番目ユニットの閾値，$\xi^{(\hat{k}+1)} = (\xi_1^{(\hat{k}+1)}, \ldots, \xi_{M_{\hat{k}+1}}^{(\hat{k}+1)})^T$ $(2 \le j_{\hat{k}} \le M_{\hat{k}})$ とする．図 2.2 は $(M_{\hat{k}}-1)$-ニューラルネットワークを表す.

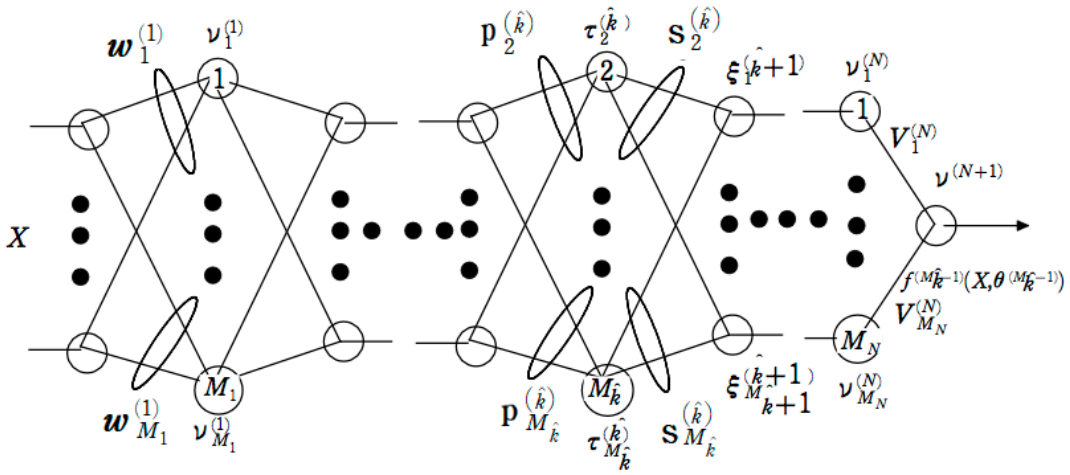


図 2.2　$(M_{\hat{k}}-1)$-ニューラルネットワーク

2.1.1　埋め込み写像の定義

$\Theta_{M_{\hat{k}-1}}$ から $\Theta_{M_{\hat{k}}}$ への 3 つの埋め込み写像を考える．図 2.3，図 2.4，図 2.5 は埋め込み写像 $\alpha_{\tilde{\mathbf{w}}}$，$\beta_{(\mathbf{u},\nu)}$，$\gamma_\lambda$ を表している ([8]).

定義 8　$(1)\tilde{\mathbf{w}} = (\mathbf{w}^T, \nu)^T \in \mathbb{R}^{M_{\hat{k}-1}+1}$ に対して,

$$\alpha_{\tilde{\mathbf{w}}} : \Theta_{M_{\hat{k}}-1} \to \Theta_{M_{\hat{k}}}$$

$$\theta^{(M_{\hat{k}}-1)} \mapsto (\ldots, \xi^{(\hat{k}+1)^T}, \mathbf{s}_1^{(\hat{k})^T} \equiv \mathbf{0}^T, \mathbf{s}_2^{(\hat{k})^T}, \ldots, \mathbf{s}_{M_{\hat{k}}}^{(\hat{k})^T}, \tilde{\mathbf{w}}^T, \tilde{\mathbf{p}}_2^{(\hat{k}^T)}, \ldots, \tilde{\mathbf{p}}_{M_{\hat{k}}}^{(\hat{k}^T)}, \ldots)^T$$

ここで,

$$\theta^{(M_{\hat{k}}-1)} = (\ldots, \xi^{(\hat{k}+1)^T}, \mathbf{s}_2^{(\hat{k})^T}, \ldots, \mathbf{s}_{M_{\hat{k}}}^{(\hat{k})^T}, \tilde{\mathbf{p}}_2^{(\hat{k}^T)}, \ldots, \tilde{\mathbf{p}}_{M_{\hat{k}}}^{(\hat{k}^T)}, \ldots)^T.$$

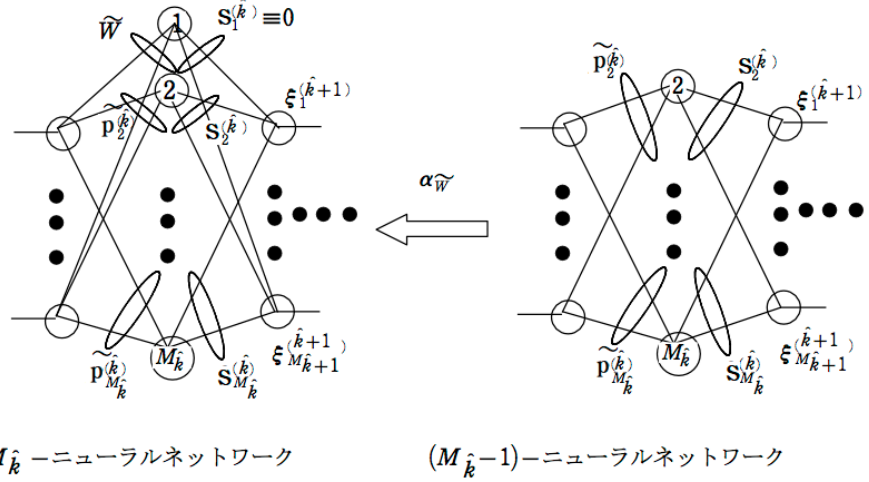


図 2.3　埋め込み写像 $\alpha_{\tilde{\mathbf{w}}}$

(2) $\mathbf{u} \in \mathbb{R}^{M_{\hat{k}+1}}$,　$\nu \in \mathbb{R}$ に対して,

$$\beta_{(\mathbf{u},\nu)} : \Theta_{M_{\hat{k}}-1} \to \Theta_{M_{\hat{k}}}.$$

$$\theta^{(M_{\hat{k}}-1)} \mapsto (\ldots, \xi^{(\hat{k}+1)^T} - \varphi(\nu)\mathbf{u}^T, \mathbf{u}^T, \mathbf{s}_2^{(\hat{k})^T}, \ldots, \mathbf{s}_{M_{\hat{k}}}^{(\hat{k})^T}, (\mathbf{0}^T, \nu), \tilde{\mathbf{p}}_2^{(\hat{k}^T)}, \ldots, \tilde{\mathbf{p}}_{M_{\hat{k}}}^{(\hat{k}^T)}, \ldots)^T$$

ここで,

$$\theta^{(M_{\hat{k}}-1)} = (\ldots, \xi^{(\hat{k}+1)^T}, \mathbf{s}_2^{(\hat{k})^T}, \ldots, \mathbf{s}_{M_{\hat{k}}}^{(\hat{k})^T}, \tilde{\mathbf{p}}_2^{(\hat{k}^T)}, \ldots, \tilde{\mathbf{p}}_{M_{\hat{k}}}^{(\hat{k}^T)}, \ldots)^T.$$

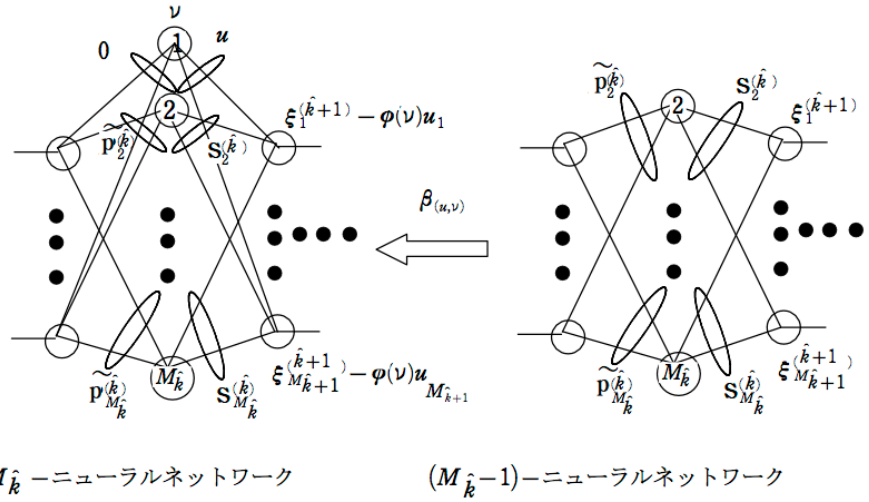


図 2.4　埋め込み写像 $\beta_{(\mathbf{u},\nu)}$

(3) $\lambda \in \mathbb{R}$ に対して, $\gamma_\lambda : \Theta_{M_{\hat{k}}-1} \to \Theta_{M_{\hat{k}}}$

$$\theta^{(M_{\hat{k}}-1)} \mapsto (\ldots, \xi^{(\hat{k}+1)^T}, \lambda \mathbf{s}_2^{(\hat{k})^T}, (1-\lambda)\mathbf{s}_2^{(\hat{k})^T}, \mathbf{s}_3^{(\hat{k})^T}, \ldots, \mathbf{s}_{M_{\hat{k}}}^{(\hat{k})^T}, \tilde{\mathbf{p}}_2^{(\hat{k}^T)}, \tilde{\mathbf{p}}_2^{(\hat{k}^T)}, \tilde{\mathbf{p}}_3^{(\hat{k}^T)}, \ldots, \tilde{\mathbf{p}}_{M_{\hat{k}}}^{(\hat{k}^T)}, \ldots)^T$$

ここで,

$$\theta^{(M_{\hat{k}}-1)} = (\ldots, \xi^{(\hat{k}+1)^T}, \mathbf{s}_2^{(\hat{k})^T}, \ldots, \mathbf{s}_{M_{\hat{k}}}^{(\hat{k})^T}, \tilde{\mathbf{p}}_2^{(\hat{k}^T)}, \ldots, \tilde{\mathbf{p}}_{M_{\hat{k}}}^{(\hat{k}^T)}, \ldots)^T.$$

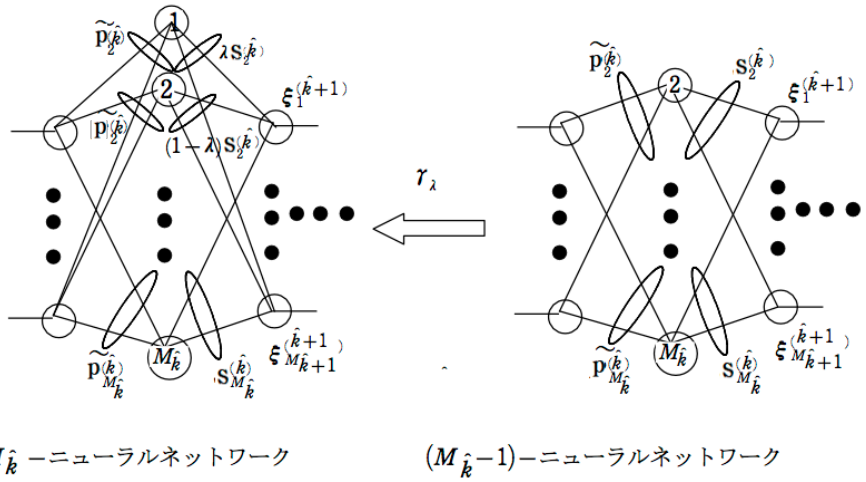


図 2.5　埋め込み写像 γ_λ

2.1.2　臨界点の埋め込み定理

定理 1 ([8])　$\theta_*^{(M_{\hat{k}}-1)} \in \Theta_{M_{\hat{k}}-1}$ が $(M_{\hat{k}}-1)$-ニューラルネットワークの誤差関数 $E_{M_{\hat{k}}-1}$ の臨界点とする.

(1) $\alpha_{\tilde{\mathbf{w}}}$ について, $\tilde{\mathbf{w}} = (\mathbf{0^T}, \nu)^\mathbf{T}$ のとき, $\theta_*^{(M_{\hat{k}})} = \alpha_{\tilde{\mathbf{w}}}(\theta_*^{(M_{\hat{k}}-1)})$ は $M_{\hat{k}}$-ニューラルネットワークの誤差関数 $E_{M_{\hat{k}}}$ の臨界点である.

(2) $\beta_{(\mathbf{u},\nu)}$ について, $\theta_*^{(M_{\hat{k}})} = \beta_{(\mathbf{0},\nu)}(\theta_*^{(M_{\hat{k}}-1)})$ は $M_{\hat{k}}$-ニューラルネットワークの誤差関数 $E_{M_{\hat{k}}}$ の臨界点である.

(3) γ_λ について, $\theta_*^{(M_{\hat{k}})} = \gamma_\lambda(\theta_*^{(M_{\hat{k}}-1)})$ は $M_{\hat{k}}$-ニューラルネットワークの誤差関数 $E_{M_{\hat{k}}}$ の臨界点である.

2.2 過剰般化

「過剰般化の現象」と「過学習の現象」は別の現象ではあるが，本研究においては，これまで「過学習」と呼ばれてきた現象については，分析の対象とはせず，過剰般化現象に焦点を絞り考察する．「過学習」については呼称と事例を示すのみとする．

負の転移について，英語学習において次の英文のような誤りが発生する．

Mr. Suzuki teached us English.

この場合，動詞 teach の過去形を正しい形の taught ではなく teached としてしまう．

このような誤りは，母語である日本語の構造からの転移であるとは考えにくい．第二言語のある規則をその規則が適用できない他の項目まで当てはめてしまったために生じた誤りであると考えるのが一般的である．このような誤りの原因を「過剰般化 (overgeneralization)」と呼び英語学習の問題点として知られている（[1]）．

2.2.1 順列・組合せへの応用

場合の数において学習順序として一般的に順列から組合せの学習を行っている．順列は「並べる」ときの概念であるが，並べるためには「選んで」から「並べる」という 2 つの操作を行う必要がある．その際，初めは「選ぶ」ことについてあまり意識せずに「並べる」ことについて学習をする．その後，組合せとして「選ぶ」概念を学習する．順列は組合せより計算が簡単であるが，一方で意味を理解するには「選んで並べる」という組合せの概念も含んでいる．

順列から組合せへと学習することの順序性によって起こる組合せの問題を順列と解釈してしまう状態を，「組合せを順列とする過剰般化」と定める．記号 ${}_n\mathrm{C}_r$ を ${}_n\mathrm{P}_r$ として計算したり「選ぶ」という言葉の意味を「並べる」と考えてしまう現象とする．ここで順列の学習をすることで過剰に組合せに影響して順列の適用範囲が広がっている．また，組合せ学習後に順列を再学習することの順序性によって起こる順列の問題を組合せと解釈してしまう状態を，「順列を組合せとする過剰般化」と定める．記号 ${}_n\mathrm{P}_r$ を ${}_n\mathrm{C}_r$ として計算したり「並べる」という言葉の意味を「選ぶ」と考えてしまう現象とする．ここで組合せの学習が過剰に順列に影響して組合せの適用範囲が広がっている．

順列と組合せの互いの理解を必要とする問題に対して，正答しても組合せの理解が不十分で順列の理解だけで考えてしまう場合は本当に互いの理解が進んでいるのか，互いの理解ができた後もどのような過剰般化の影響が再び起こりやすいか，また，過剰般化が起こ

り正答できなかった場合も，順列の再学習をして深い理解をするためには，組合せの初期段階での順列を組合せとする過剰般化は学習を進める上で必要な過程であることについて第 9 章に後述する．

2.2.2　分類測度を用いた可視化

以下は [20] で明らかにしたことに基づく．

3 種類のあやめのデータ (アイリスデータ) に対して，1,2 組を学習 (分類) したニューラルネットワークに対して, 3 組のデータの学習 (分類) を行う．1,2 組から 3 組へと学習することの順序性によって起こる 3 組を 2 組と分類してしまう現象を「3 組を 2 組とする過剰般化」と定める．ここで 2 組の学習が過剰に 3 組に影響して 2 組の適用範囲が広がっている．また 3 組から 2 組へと学習することの順序性によって起こる 2 組を 3 組と分類してしまう現象を「2 組を 3 組とする過剰般化」と定める, ここで 3 組の学習が過剰に 2 組に影響して 3 組の適用範囲が広がっている．

入力が 4 次元ベクトル, 出力が組である 4 層ニューラルネットワーク（線形層，tanh，線形層，ソフトマックス層）を作成するため，Mathematica を用いて次のように入力すると，以下の図 2.6 の左側に示す．

```
parameterNet =NetChain[{LinearLayer[5], ElementwiseLayer[Tanh],
                        LinearLayer[3], SoftmaxLayer[]}, "Input" -> 4, "Output" -> dec]
```

入力 (4 次元ベクトルをニューラルネットワークに通した出力値) と, 対象 (組) のクロスエントロピー損失を求める．Mathematica を用いて次のように入力してネットグラフを作成すると，以下の図 2.6 の右側に表示される．

```
trainingNet =NetGraph[<|"net" -> net, "loss" -> CrossEntropyLossLayer["Index"]|>,
                      {NetPort["Input"] -> "net" -> NetPort["loss", "Input"],
                       NetPort["Target"] -> NetPort["loss", "Target"]}]
```

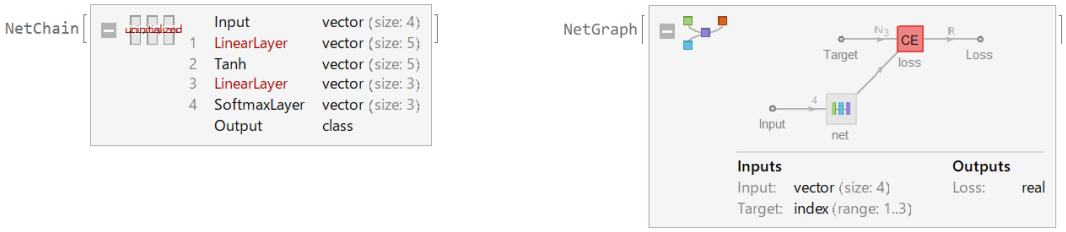


図 2.6　ニューラルネットワーク，ネットグラフ

2.2.3　アイリスデータ (1,2 組) の学習

バッチサイズを 35 として，最大 500 ラウンド学習する中でテストデータ (1,2 組) の損失が最小となるラウンド数で学習を終えるように，訓練データ (1,2 組) を学習させるために次のように入力する:

$$\texttt{results} = \texttt{NetTrain}[\texttt{trainingNet}, \texttt{train}[1,2], \texttt{All}, \texttt{ValidationSet} -> \texttt{test}[1,2],$$
$$\texttt{BatchSize} -> 35, \texttt{MaxTrainingRounds} -> 500]$$

学習結果の要約と, 損失の変化が図 2.7 の左側に示す.

学習後のネットを trainednet と定めて，次のように入力しテストデータ (1,2 組) の分類測度を求める.

$$\texttt{measurements} = \texttt{ClassifierMeasurements}[\texttt{trainednet}, \texttt{testset}[1,2]]$$

正確さは 1 であり分類測度と分類表が図 2.7 の右側に示す.

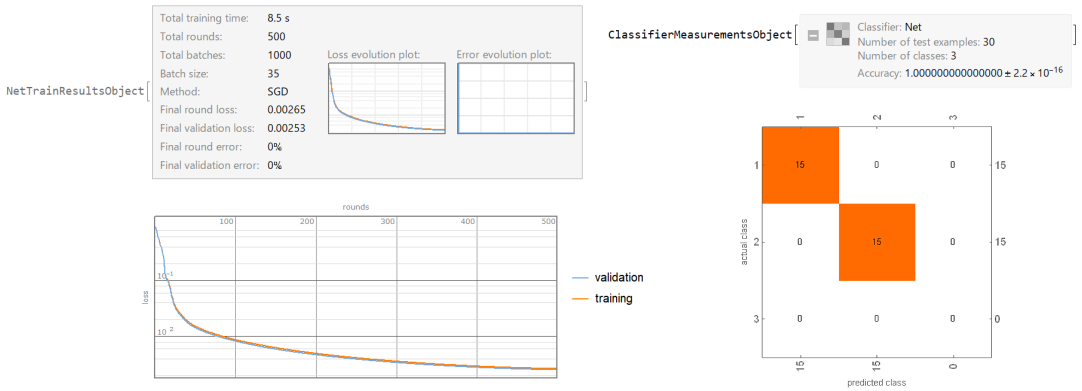


図 2.7　アイリスデータ (1,2 組) の学習，分類

2.2.4　アイリスデータ (1,2,3 組) の学習

学習後のニューラルネットワークにおいて, バッチサイズを 35 として訓練データ (1,2,3 組) の学習をさせる．テストデータ (3 組) の損失が最小となるラウンド数で学習を終える.

ラウンド数が 8 のとき次のように入力する:

$$\texttt{results1} = \texttt{NetTrain}[\texttt{results}[\text{"TrainedNet"}], \texttt{train}[1,2,3], \texttt{All},$$
$$\texttt{ValidationSet} -> \texttt{test}[3], \texttt{BatchSize} -> 35, \texttt{MaxTrainingRounds} -> 8]$$

損失が最小であるラウンド数は 8 であり，学習結果の要約と, 損失の変化が図 2.8 の左側に示す.

学習後のネットを trainednet1 と定めて，次のように入力しテストデータ (1,2,3 組) の分類測度を求める.

$$\texttt{measurements} = \texttt{ClassifierMeasurements}[\texttt{trainednet1}, \texttt{testset}[1, 2, 3]]$$

正確さは 0.666667 であり分類測度と分類表が図 2.8 の右側に示す.

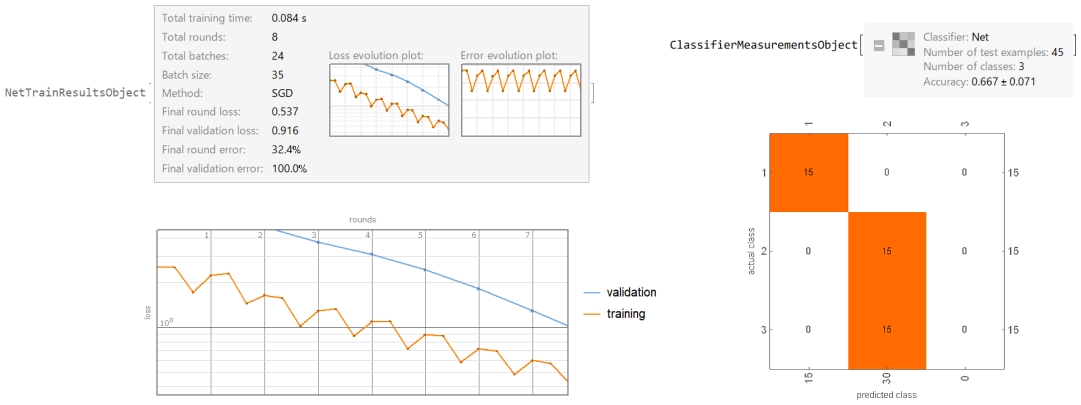


図 2.8　アイリスデータ (1,2,3 組) の学習，分類

ラウンド数が 30 のとき次のように入力する:

$$\texttt{results2} = \texttt{NetTrain}[\texttt{results}[\text{"}\texttt{TrainedNet}\text{"}], \texttt{train}[1, 2, 3], \texttt{All},$$
$$\texttt{ValidationSet} -> \texttt{test}[3], \texttt{BatchSize} -> 35, \texttt{MaxTrainingRounds} -> 30]$$

損失が最小であるラウンド数は 13 であり，学習結果の要約と, 損失の変化が図 2.9 の左側に示す.

学習後のネットを trainednet2 と定めて，次のように入力しテストデータ (1,2,3 組) の分類測度を求める.

$$\texttt{measurements} = \texttt{ClassifierMeasurements}[\texttt{trainednet2}, \texttt{testset}[1, 2, 3]]$$

正確さは 0.666667 であり分類測度と分類表が図 2.9 の右側に示す.

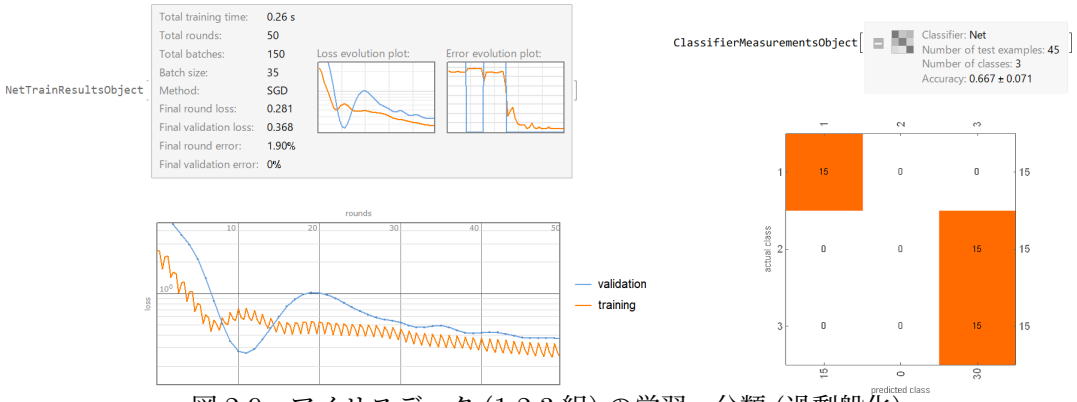


図 2.9　アイリスデータ (1,2,3 組) の学習，分類 (過剰般化)

学習後のニューラルネットワークにおいて, バッチサイズを 35 として訓練データ (1,2,3 組) の学習をさせる．テストデータ (2,3 組) の損失が最小となるラウンド数で学習を終える．

ラウンド数が 100 のとき次のように入力する:

$$\texttt{results3} = \texttt{NetTrain}[\texttt{results2}[\text{"}\texttt{TrainedNet}\text{"}], \texttt{train}[1,2,3], \texttt{All},$$
$$\texttt{ValidationSet} -> \texttt{test}[2,3], \texttt{BatchSize} -> 35, \texttt{MaxTrainingRounds} -> 100]$$

学習結果の要約と, 損失の変化が図 2.10 の左側に示す.

学習後のネットを trainednet3 と定めて，次のように入力しテストデータ (1,2,3 組) の分類測度を求める.

$$\texttt{measurements} = \texttt{ClassifierMeasurements}[\texttt{trainednet3}, \texttt{testset}[1,2,3]]$$

正確さは 1 であり分類測度と分類表が図 2.10 の右側に示す.

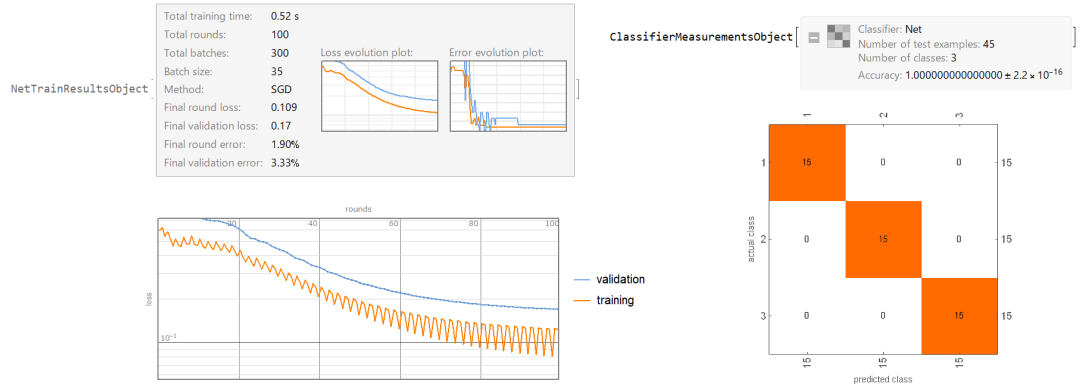


図 2.10　アイリスデータ (1,2,3 組) の学習，分類

第 3 章

情報科学における学習の捉え方

渡辺 ([21]) による特異点解消定理を用いて，学習・汎化損失について述べ，カルバック情報量や学習係数を計算する．数式処理システム Mathematica を用いて，中間ユニット数を変えて過学習を起こす場合と起こさない場合について分析を行う．起こらない場合は RMS 重みの分析を行う．

3.1　学習理論

3.1.1　学習システム

渡辺 ([21]) は「人間の機能を捉える」問題について関数を用いて考察している．しかし，関数が求まったとしても「何ができたら人間の機能といえるのか」はまだ定義されていないままである．「愛する人の幸福を願う人工知能」が作られたとき，「愛する人の幸福を【本当に】願っているのか」については現在まで不明なままであり図 3.1 で表す．

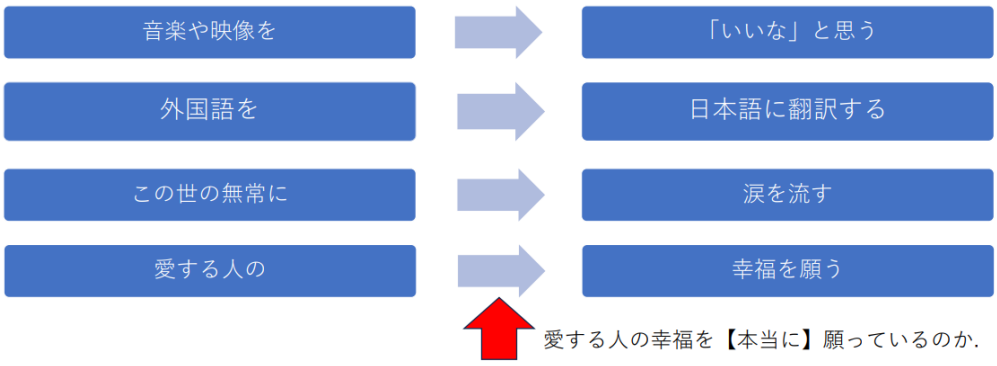


図 3.1　「人間の機能を捉える」問題

情報科学における学習理論では，ニューラルネットワークのパラメータを更新して，誤差関数を最小にするパラメータを求めることを学習という．

3.1.2 真の分布と学習モデル

以下の定義 9 から定義 23，定理 2 から定理 6，命題 1,2，系 1,2（[3]）を用いる．

定義 9 (関数近似モデル) 入力 X が従う確率分布 $q(x)$ と，パラメータ w をもつ $\mathbb{R}^1$ から $\mathbb{R}^1$ への関数 $f(x,w)$ に対して，平均 0，分散 1 の $\mathbb{R}^N$ 上の確率変数 Z に対して，$\mathbb{R}^N$ 上の確率変数 Y を $Y = f(x,w) + Z$ と定め，関数近似モデルと言う．

定義 10 (学習モデル) $\sigma \in \mathbb{R}^1$ を標準偏差とする． $x, y \in \mathbb{R}^1$， $w \in \mathbb{R}^d$，$\mathbb{R}^1$ 値関数 $f(x,\ w)$ と確率密度関数 $p(y|x,w)$ に対して次が成り立つ．

$$p\left(y|x,w\right) := \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{|y - f\left(x,w\right)|^2}{2\sigma^2}\right).$$

確率密度関数 $p\left(x\right)$ に従う入力 X，確率密度関数 $p(y|x,w)$ に対して，確率密度関数 $p\left(x,y|w\right)$ を次で定める．

$$p\left(x,y|w\right) = p\left(x\right)p\left(y|x,w\right).$$

そのとき条件付き確率密度関数を学習モデルと呼び，次で定める．

$$p\left(x,y|w\right) = \frac{p\left(x\right)}{\sqrt{2\pi}\sigma} \exp\left(-\frac{|y - f\left(x,w\right)|^2}{2\sigma^2}\right).$$

定義 11 (真の分布) $\sigma \in \mathbb{R}^1$ を標準偏差とする． $x, y \in \mathbb{R}^1$， $w_0 \in \mathbb{R}^d$，$\mathbb{R}^1$ 値関数 $f(x,\ w_0)$, と確率密度関数 $p(x)$ に対して条件付き確率密度関数を真の分布と呼び，次で定める．

$$q\left(x,y\right) := p\left(x,y|w_0\right) = \frac{p\left(x\right)}{\sqrt{2\pi}\sigma} \exp\left(-\frac{|y - f\left(x,w_0\right)|^2}{2\sigma^2}\right).$$

定義 12 (予測分布)　例 X^n が与えられたとき,
パラメータ w の確率密度関数

$$p(w|X^n) := \frac{1}{Z(X^n)}\varphi(w)\prod_{i=1}^{n} p(X_i|w),$$

を事後分布という.
ここで $Z(X^n)$ は正規化定数

$$Z(X^n) := \int \varphi(w)\prod_{i=1}^{n} p(X_i|w)\,dw,$$

であり, 分配関数という.

このとき, x の確率密度関数を

$$p(x|X^n) := \int p(x|w)p(w|X^n)\,dw,$$

と定め, 予測分布という.

真の分布 $q(x)$ は予測分布 $p(x|X^n)$ に近いであろうと推測する. この推測のことを学習という.

3.1.3　カルバック情報量

定義 13 (カルバック情報量)　任意の確率密度関数 $q(x), p(x) > 0$, 開集合 $A \subset \mathbb{R}^N$ に対して, カルバック情報量又は相対エントロピーを次で定める.

$$K(q||p) := \int q(x)\ \log\frac{q(x)}{p(x)}dx.$$

命題 1　カルバック情報量は次の性質を満たす.
(1) 任意の $q(x), p(x)$ に対して, $K(q||p) \geq 0$.
(2) $K(q||p) = 0 \Leftrightarrow q(x) = p(x)$.

カルバック情報量を用いて真の分布 $q(x)$ と予測分布 $p(x|X^n)$ の違いを測る.

3.2 特異点の定義

3.2.1 臨界点と特異点

定義 14 (臨界点) U を $\mathbb{R}^d$ の開集合，$f : U \longrightarrow \mathbb{R}^1$ を C^1 級の関数とする．f の臨界点を次を満たす点 $x^* \in U$ として定める．

$$\nabla f(x^*) = \left(\frac{\partial f}{\partial x_1}(x^*), \frac{\partial f}{\partial x_2}(x^*), \cdots, \frac{\partial f}{\partial x_d}(x^*) \right) = \mathbf{0}.$$

定義 15 (特異点) A を $\mathbb{R}^d$ の空でない開集合とする．
(1) A 上の点 P が非特異であるとは，次の条件を満たす点のことである．
開集合 $U(\mathrm{P} \in U), V \subset \mathbb{R}^d$ と C^ω 級微分同相写像 $f : U \longrightarrow V$ に対して

$$f(A \cap U) = \{(x_1, x_2, \cdots, x_n, 0, 0, \cdots, 0) | \ x_i \in \mathbb{R} \} \cap V.$$

を満たすものが存在する．ここで，n は自然数である．A の全ての点を非特異点といい，A を非特異点集合と呼ぶ.
(2) A の点 P は非特異点ではないとする．このとき特異点である，または集合 A の特異点集合であるという． A の特異点集合は A の特異点部分と呼び, 次で表される．

$$Sing(A) = \{\mathrm{P} \in A | \mathrm{P} \text{ is a singularity of } A \}.$$

定義 16 (汎化誤差) 真の分布 $q(x)$ と予測分布 $p(x|X^n)$ に対して汎化誤差 $G(X^n)$ を次で定める ．

$$G(X^n) := \int q(x) \log \frac{q(x)}{p(x|X^n)} dx.$$

次に対数尤度比関数を次で定める．

$$K_n(w) := \frac{1}{n} \sum_{i=1}^{n} \log \frac{q(X_i)}{p(X_i|w)}.$$

事後分布は

$$p\left(w|X^n\right) = \frac{1}{Z\left(X^n\right)}\varphi\left(w\right)\prod_{i=1}^{n} p\left(X_i|w\right).$$

次のように書き表される.

$$p\left(w|X^n\right) = \frac{1}{Z_0\left(X^n\right)}\exp\left(-nK_n\left(w\right)\right)\varphi\left(w\right).$$

ここで

$$Z_0\left(X^n\right) := \int \exp\left(-nK_n\left(w\right)\right)\varphi\left(w\right)dw.$$

定義 17 (確率的複雑さ) $Z_0\left(X^n\right)$ に対して確率的複雑さを次で定める.

$$F\left(X^n\right) := -\log Z_0\left(X^n\right).$$

真の分布 $q\left(x\right)$ 学習モデル $p(x|w)$ に含まれている場合を考える.
次のように定めると

$$W_0 := \{w \in \mathbb{R}^d \,|q\left(x\right) = p\left(x|w\right)\} = \{w \in \mathbb{R}^d \mid K(w) = 0\}.$$

$W_0 \neq \emptyset$. が成り立つ.

定義 18 (ゼータ関数) カルバック情報量 $K\left(w\right)$, $z \in \mathbb{C}$, 事前分布 $\varphi\left(w\right)$ に対してゼータ関数を次で定める.

$$\zeta\left(z\right) := \int K\left(w\right)^z \varphi\left(w\right)dw.$$

3.2.2 特異点解消定理

定義 19 (解析多様体) ハウスドルフ空間 M に対して, M を被覆する開集合の族 $\{U\}$ が存在し, 任意の U に対して同相写像 $\phi : U \ \rightarrow \phi(U) \subset \mathbb{R}^d$ が存在するとする. このとき, $\{(U, \phi)\}$ を M の座標近傍系といい, M を d 次元多様体という.

M の 2 つの座標近傍系 $\{(U_1, \phi_1)\}$, $\{(U_2, \phi_2)\}$ が与えられて，$U^* \equiv U_1 \cap U_2 \neq \emptyset$ であるとき，2 つの関数

$$\phi_1(U^*) \ni x \mapsto \phi_2(\phi_1^{-1}(x)) \in \phi_2(U^*)$$

$$\phi_2(U^*) \ni x \mapsto \phi_1(\phi_2^{-1}(x)) \in \phi_1(U^*)$$

がともに，r 回連続微分できるならば，M を C^r 級微分多様体という．この 2 つの関数が，ともに解析関数 *(C^ω 級)* であるとき M を解析多様体という．

定義 20 (既約な代数的集合) 代数的集合 V *(*定義 35 参照*)* が既約 *(irreducible)* であるとは $V = V_1 \cup V_2$ ならば $V = V_1$ または $V = V_2$ が成り立つ．

定義 21 (代数的集合の次元) 既約な代数的集合 V 上で *0* になる多項式全体の集合 $I(V)$ *(*定義 36 参照*)* を考える．多項式 $f_1(x), f_2(x), \cdots, f_r(x)$ が $I(V)$ の生成元であるとする．

$$I(V) = \langle f_1, f_2, \cdots, f_r \rangle.$$

ヤコビ行列

$$J(x) = \begin{pmatrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_1(x)}{\partial x_d} \\ \cdots & \cdots & \cdots \\ \frac{\partial f_r(x)}{\partial x_1} & \cdots & \frac{\partial f_r(x)}{\partial x_d} \end{pmatrix}$$

の V 上での $J(x)$ のランクの最大値を d_0 とする．このとき，$d - d_0$ を既約な代数的集合 V の次元という．

定理 2 (非特異, 特異点) V 上の点 x が非特異であることと $J(x)$ のランクが d_0 になることは同値である．また x が特異点であることと $J(x)$ のランクが d_0 よりも小さくなることとは同値である．

定義 22 (ブローアップ) d 次元のユークリッド空間 $\mathbb{R}^d$ を考える．自然数 r が $2 \leq r \leq d$ を満たすとする．$d - r$ 次元の非特異代数的集合

$$V = \{x \in \mathbb{R}^d | x_1 = x_2 = \cdots = x_r = 0\}.$$

を中心とする $\mathbb{R}^d$ の中の代数的集合 W のブローアップを，直積集合 $\mathbb{R}^d \times \mathbb{P}^{r-1}$ の部分集合として次で定める．

$$B_V(W) \equiv \overline{\{(x,(x_1:x_2:\cdots:x_r));x \in W \setminus V\}}.$$

定理 3 (特異点解消定理) 任意の空でない代数的集合 $W \subset \mathbb{R}^d$ について, 次のような代数的真部分集合 V が存在する．
(1) 代数的集合 V は W の特異点集合に含まれる．
(2) $B_V(W)$ が $\mathbb{R}^d \times \mathbb{P}^{r-1}$ の中の非特異な代数的集合になる．
(3) (2) の $B_V(W)$ は,「特異点集合に含まれる非特異集合を中心とするブローアップ」の有限回の繰り返しによってつくりだすことができる．

定義 23 (正規交差) $\mathbb{R}^d$ の中のある開集合 U から $\mathbb{R}$ への実解析的な関数 $f(x)$ が $x^* = (x_1^*, x_2^*, \cdots, x_d^*)$ において正規交差であるとは，x^* の近傍で x^* を原点とする座標系 $y = (y_1, y_2, \cdots, y_d)$ と非負の自然数 $k_1, k_2, \cdots, k_d$ が存在して

$$f(y) = a\,(y)\, y_1^{k_1} y_2^{k_2} \ldots {y_d}^{k_d}$$

と書けることである．ここで $a\,(y)$ は U 上で $|a\,(y)\,| > 0$ を満たす実解析関数である．

3.2.3 確率的複雑さと汎化誤差の漸近展開

渡辺 ([3]) は漸近展開について以下のことを明らかにした．

定理 4 ゼータ関数は特異点解消写像 $w = g\,(u)$ を用いて次で書き直される．

$$\zeta\,(z) = \int K\,(g\,(u))^z\, \varphi\,(g\,(u))\, |g'\,(u)|\, du.$$

(1) ゼータ関数は $Re\,(z) > 0$ において正則である．
(2) $\zeta\,(z)$ は複素数平面 C の正則関数に含まれ解析的である．その極は非負の実数であり, 有理数である．大きいものから小さいものへと順番に

$$(-\lambda_1)\,, (-\lambda_2)\,, \ldots,$$

とする．それぞれの極 $(-\lambda_k)$ の位数を m_k とする．

このときゼータ関数 $\zeta(z)$ は次のようにローラン展開される.

$$\zeta(z) = \zeta_0(z) + \sum_{k=1}^{\infty} \sum_{m=1}^{m_k} \frac{c_{km}}{(z+\lambda_k)^m}.$$

ここで $\zeta_0(z)$ 正則関数で $\{c_{km}\}$ は係数である.

定理 5 確率的複雑さは

$$F(X^n) = -\log \int \exp(-nK_n(w))\varphi(w)\, dw,$$

次のように漸近展開される.

$$F(X^n) = \lambda_1 \log n - (m_1 - 1)\log(\log n) + R(X^n).$$

確率変数 $R(X^n)$ が存在して, R に 法則収束する.

$$R(X^n) \to R$$

$$\lim_{n\to\infty} E_{X^n}[R(X^n)] = E[R].$$

系 1 $F(n)$ を $F(n) = E_{X^n}[F(X^n)]$ と表す. $F(n)$ は次の条件を満足する.

$$F(n) = \lambda_1 \log n - (m_1 - 1)\log(\log n) + O(1).$$

定理 6 汎化誤差は

$$G(X^n) = \int q(x) \log \frac{q(x)}{p(x|X^n)} dx$$

次のように漸近展開される.

$$G(X^n) = \frac{G_0(X^n)}{n}.$$

確率変数 $G_0(X^n)$ が存在して, G_0に 法則収束する.

$$nG_0(X^n) \to G_0$$

$$\lim_{n\to\infty} E_{X^n}[G_0(X^n)] = E[G_0] = \lambda_1.$$

系 2 $G(n)$ を $G(n) = E_{X^n}[G(X^n)]$ と表す．$G(n)$ は次の条件を満足する．

$$G(n) = \frac{\lambda_1}{n} + O\left(\frac{1}{n}\right).$$

3.2.4 カルバック情報量，学習係数の計算 (1 つの被覆における)

命題 2 学習モデル $p(y|x,w)\,q(x)$ と真の分布 $q(y|x)\,q(x)$ がそれぞれ次のように与えられている．

$$p(y|x,w) = \exp\left(-\frac{1}{2}(y - f(x,w))^2\right),$$

$$q(y|x) = \exp\left(-\frac{1}{2}(y - f_0(x))^2\right).$$

そのときカルバック情報量は次で表される．

$$K(w) = \frac{1}{2}\int (f(x,w) - f_0(x))^2\, q(x)\, dx.$$

学習モデルとして中間ユニット数が 2 である 3 層ニューラルネットワーク（第 1 層から第 2 層への重みを b_1 と b_2，第 2 層から第 3 層への重みを a_1 と a_2，バイアスをなしとする．）が中間ユニット数が 0 である真のモデルを実現する場合を考える．

3 層ニューラルネットワークのカルバック情報量について計算をする．

$$f = (a_1b_1 + a_2b_2)^2 + \left(a_1b_1^3 + a_2b_2^3\right)^2.$$

学習モデルに特異点解消定理を用いて分析する．

(1) 中心 $V(b_1,b_2) \subset V(f)$ でブローアップする．$b_1 = b_{11}b_2$ によって

$$f = (a_1b_{11}b_2 + a_2b_2)^2 + \left(a_1b_{11}^3b_2^3 + a_2b_2^3\right)^2$$

$$= b_2^2(a_1b_{11} + a_2)^2 + b_2^4\left(a_1b_{11}^3b_2^3 + a_2\right)^2.$$

f は (b_1,b_2) の対称性より $b_2 = b_{21}b_1$ する必要はない．

(2) 座標変換 $a_{21} = a_1b_{11} + a_2$ は解析同型でヤコビ行列の次元は 1 である．

$$f = b_2^2\{a_{21}^2 + b_2^4\left(a_1b_{11}^3 + a_{21} - a_1b_{11}\right)^2\}.$$

(3) 2 回目は $V(a_{21}, b_2) \subset V(f)$ を中心とするブローアップを行う．第 1 座標において $b_2 = a_{21}b_{21}$ によって次が成り立つ．

$$f = a_{21}^4b_{21}^2\{1 + b_{21}^4 + b_2^4\left(a_1b_{11}^3 + a_{21} - a_1b_{11}\right)^2\}.$$

これは正規交差である．第 2 座標 $a_{21} = a_{22}b_2$ によって次が成り立つ．

$$f = b_2^4\{a_{22}^2 + b_2^2\left(a_1b_{11}^3 + a_{22}b_2 - a_1b_{11}\right)^2\}.$$

これは正規交差ではない．

(4) 3 回目は $V(a_{22}, b_2) \subset V(f)$ を中心とするブローアップを行う．第 1 座標において $b_2 = a_{22}b_{21}$ によって次が成り立つ．

$$\begin{aligned} f &= a_{22}^4b_{21}^4\{a_{22}^2 + a_{22}^2b_{21}^2\left(a_1b_{11}^3 + a_{22}a_{22}b_{21} - a_1b_{11}\right)^2\} \\ &= a_{22}^6b_{21}^4\{1 + b_{21}^2\left(a_1b_{11}^3 + a_{22}a_{22}b_{21} - a_1b_{11}\right)^2\}. \end{aligned}$$

これは正規交差である．第 2 座標 $a_{22} = a_{23}b_2$ によって次が成り立つ．

$$\begin{aligned} f &= b_2^4\{a_{23}^2b_2^2 + b_2^2\left(a_1b_{11}^3 + a_{23}b_2^2 - a_1b_{11}\right)^2\} \\ &= b_2^6\{a_{23}^2 + \left(a_1b_{11}^3 + a_{23}b_2^2 - a_1b_{11}\right)^2\}. \end{aligned}$$

これは正規交差ではない．

(5) 4 回目は $V(a_{23}, a_1) \subset V(f)$ を中心とするブローアップを行う．第 1 座標において $a_1 = a_{23}a_{12}$, によって次が成り立つ．

$$f = a_{23}^2b_2^6\{1 + \left(a_{12}b_{11}^3 + b_2^2 - a_{12}b_{11}\right)^2\}.$$

これは正規交差である．第 2 座標 $a_{23} = a_1a_{24}$ によって次が成り立つ．

$$\begin{aligned} f &= b_2^6\{a_1^2a_{24}^2 + \left(a_1b_{11}^3 + a_1a_{24}b_2^2 - a_1b_{11}\right)^2\} \\ &= a_1^2b_2^6\{a_{24}^2 + \left(b_{11}^3 + a_{24}b_2^2 - b_{11}\right)^2\}. \end{aligned}$$

これは正規交差ではない．

(6) 5 回目は $V(a_{24}, b_{11}) \subset V(f)$ を中心とするブローアップを行う．第 1 座標において $b_{11} = a_{24}b_{12}$, によって次が成り立つ．

$$f = a_1^2b_2^6a_{24}^2\{1 + \left(a_{24}^2b_{12}^3 + b_2^2 - b_{12}\right)^2\}.$$

これは正規交差である．第 2 座標 $a_{24}=b_{11}a_{25}$ によって次が成り立つ．

$$f=a_1^2b_2^6\{b_{11}^2a_{25}^2+\left(b_{11}^3+b_{11}a_{25}b_2^2-b_{11}\right)^2\}$$
$$=a_1^2b_2^6b_{11}^2\{a_{25}^2+\left(b_{11}^2+a_{25}b_2^2-1\right)^2\}.$$

これは正規交差ではない．

(7) 最後は $V\left(a_{25},b_{11}-1\right)\subset V\left(f\right)$ を中心とするブローアップを行う．第 1 座標において $b_{11}-1=a_{25}b_{12},$ によって次が成り立つ．

$$f=a_1^2b_2^6b_{11}^2\{a_{25}^2+\left(a_{25}^2b_{12}^2+2a_{25}b_{12}+1+a_{25}b_2^2-1\right)^2\}$$
$$=a_1^2b_2^6\left(a_{25}b_{12}+1\right)^2a_{25}^2\{1+\left(a_{25}b_{12}^2+2b_{12}+b_2^2\right)^2\}.$$

これは正規交差である．第 2 座標 $a_{25}=\left(b_{11}-1\right)a_{26}$ によって次が成り立つ．

$$f=a_1^2b_2^6b_{11}^2\{\left(b_{11}-1\right)^2a_{26}^2+\left(b_{11}^2+\left(b_{11}-1\right)a_{26}b_2^2-1\right)^2\}$$
$$=a_1^2b_2^6b_{11}^2\left(b_{11}-1\right)^2\{a_{26}^2+\left(b_{11}+1+a_{26}b_2^2\right)^2\}.$$

これは正規交差である．最後の座標は次で表される．

$$a_1=a_1,\ b_1=b_{11}b_2,$$
$$a_2=a_{21}-a_1b_{11}=a_{22}b_1-a_1b_{11}=a_{23}b_2b_1-a_1b_{11}=a_{24}a_1b_2b_1-a_1b_{11}$$
$$=a_{25}b_{11}a_1b_2b_1-a_1b_{11}=a_{26}\left(b_{11}-1\right)b_{11}a_1b_2b_{11}b_2-a_1b_{11}$$
$$=a_1b_{11}\{\left(b_{11}-1\right)a_{26}b_{11}b_2^2-1\},\ b_2=b_2,$$

ヤコビ行列は次で表される．

$$\left|g'\right|=\begin{vmatrix}1&0&0&0\\0&b_2&0&b_{11}\\m_{31}&m_{32}&m_{33}&m_{34}\\0&0&0&1\end{vmatrix}=\left|a_1b_{11}^2\left(b_{11}-1\right)b_2^3\right|.$$

ここで

$$m_{31}:=b_{11}\left(b_{11}-1\right)a_{26}b_{11}b_2^2,\ m_{32}:=a_1b_{11}^2\left(b_{11}-1\right)b_2^3,$$
$$m_{33}:=a_1b_{11}\left(b_{11}-1\right)b_{11}b_2^2,\ m_{34}:=2a_1b_{11}\left(b_{11}-1\right)a_{26}b_{11}b_2.$$

$u=\left(a_1,b_{11},a_{26},b_2\right)$ に対して方程式は次のように表される．

$$g_1=a_1b_{11}\left(b_{11}-1\right)^2b_2^3a_{26}^2b_{11},$$

$$g_2 = a_1 b_{11} \left(b_{11} - 1\right)^2 b_2^3 \{(b_{11} + 1) + b_{11} a_{26}^2 b_2^2\},$$

$$g_3 = a_1 b_{11} \left(b_{11} - 1\right)^2 b_2^3 \{\left(b_{11}^3 + b_{11}^2 + b_{11} + 1\right) + b_{11} a_{26}^2 b_2^2\} b_2^2.$$

したがって

$$f\left(x, g\left(u\right)\right) = a_1 b_{11} \left(b_{11} - 1\right)^2 b_2^3 a\left(x, u\right).$$

ここで

$$a\left(x, u\right) = \frac{12 B_2}{2!} b_{11} a_{26} x + \frac{240 B_4}{4!} \{(b_{11} + 1) + b_{11} a_{26}^2 b_2^2\} x^3 +$$

$$\sum_{k=3}^{\infty} \frac{2^{2k+2} \left(2^{2k+2} - 1\right) B_{2k+2} x^{2k+1} b_2^{2(k-2)}}{(2k+2)!} \{b_{11}^{2k-1} + b_{11}^{2k-2} \cdots + 1 + b_{11} a_{26}^2 b_2^2\}.$$

したがって次を得る.

$$K\left(g\left(u\right)\right) = \frac{a_1^2 b_2^6 b_{11}^2 \left(b_{11} - 1\right)^2}{2} \int a\left(x, w\right)^2 q\left(x\right) dx.$$

ヤコビ行列 $w = g\left(u\right)$ は次で表される.

$$\left|g'\right| = \left|a_1 b_{11}^2 \left(b_{11} - 1\right) b_2^3\right|.$$

事前分布 $\varphi\left(w\right)$ はコンパクトなサポートを持つ.

ゼータ関数は次で表される.

$$\zeta\left(z\right) = \int \{a_1^2 b_2^6 b_{11}^2 \left(b_{11} - 1\right)^2\}^z \left|a_1 b_{11}^2 \left(b_{11} - 1\right) b_2^3\right| \varphi\left(g\left(u\right)\right) da_1 db_{11} da_{26} db_2.$$

ゼータ関数の最大の極は $\lambda = \frac{3}{2}$ で位数は 1 である.

先頭係数が $\frac{3}{2}$ に等しいことが分かり，確率的複雑さと汎化誤差は次で表される.

$$F\left(n\right) = \frac{3}{2} \log n + O\left(1\right).$$

$$G\left(n\right) = \frac{3}{2n} + O\left(\frac{1}{n}\right).$$

3.3　過学習の分析

情報科学における過学習について分析を行う．関数近似する際に起こる過学習がよく知られており，図 3.2 に表す．

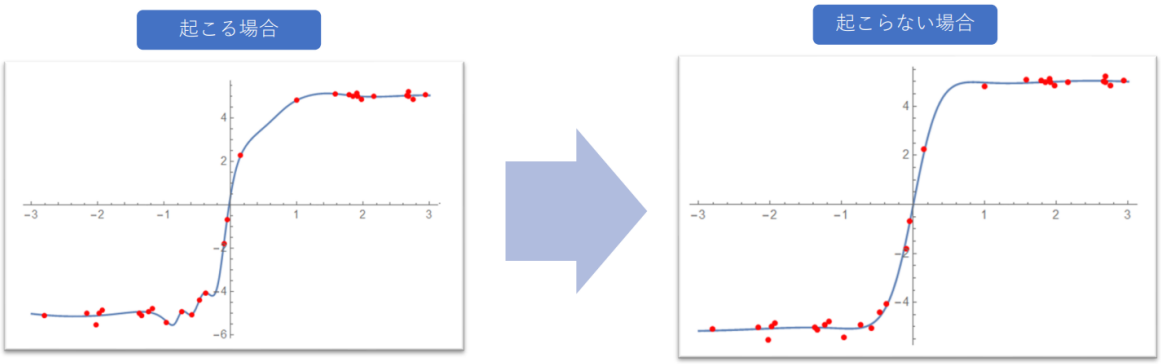


図 3.2　過学習の関数近似による分析

渡辺 ([6]) による学習・汎化損失の数式による定義を述べる．学習・汎化損失については第 6 章に，数学教育における応用については今後の課題として第 10 章に後述する．

3.3.1　経験損失と汎化損失の関係

本研究では以下の定義 24 から定義 28，定理 7 に示す手法を用いて過学習を定式化する ([6])．

定義 24 (損失関数)　真の分布を推測するため，非負の実数列 $\{a_n\}$ に対して損失関数を次のように定める．

$$R_n(w) := -\sum_{i=1}^{n} \log p(X_i|w) - a_n \log \varphi(w).$$

定義 25 (正規化された損失関数)　経験エントロピー $S_n := -\frac{1}{n}\sum_{i=1}^{n} \log q(X_i)$ を用いて $R_n(w) = R_n^0(w) + nS_n$ と表されるため，対数尤度比関数 $f(x,w) := \log \frac{q(x)}{p(x|w)}$ に対し

て正規化された損失関数を次のように定める．

$$R_n^0(w) := \sum_{i=1}^{n} f(X_i, w) - a_n \log \varphi(w).$$

ここで $R_n^0(w)$ を最小にするパラメータを $\hat{w}$ を求め，$p(x|\hat{w})$ を推測結果の確率密度関数とする．推測法として $a_n = 0$ のとき $\hat{w}$ を最尤推定法，$a_n = 1$ のとき事後確率最大化推定法という．

定義 26 (経験損失，汎化損失)　このとき汎化損失 $K(\hat{w})$ と経験損失 $K_n(\hat{w})$ を次で定める．

$$R_g = K(\hat{w}) := \int q(x) \log \frac{q(x)}{p(x|\hat{w})} dx,\ R_t = K_n(\hat{w}) := \frac{1}{n} \sum_{i=1}^{n} \log \frac{q(X_i)}{p(X_i|\hat{w})}.$$

事前分布のサポート $\mathrm{supp}(\varphi)$ はコンパクト，対数尤度比関数 $f(x, w)$ は相対的に有限な分散をもつとする．真の分布を実現するパラメータの集合 $W_0 := \{w \in \mathbb{R}^d | K(w) = 0\}$ が特異点を持つとき，$w_0 \in W_0$ として w_0 の近傍で特異点解消定理を適用させると，解析多様体 M と解析写像 $g : M \to W$ が存在して $w = g(u)$ となる．このとき自然数 $k_1, k_2, \cdots, k_r$ に対して次で表される．

$$K(g(u)) = u_1^{2k_1} u_2^{2k_2} \dots u_r^{2k_r},\ (1 \leq r \leq d).$$

また $\int a(x, u) q(x) dx = u_1^{k_1} u_2^{k_2} \dots u_r^{k_r}$ を満たす解析関数 $a(x, u)$ が存在して $f(x, g(u)) = a(x, u) u_1^{k_1} u_2^{k_2} \dots u_i^{k_i}$ と表される．このとき経験誤差関数 $K_n(g(u))$ は確率過程 $\xi_n(u) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{u^k - a(X_i, u)\}$ を用いて次で表される．

$$nK_n(g(u)) = nu^{2k} - \sqrt{n} u^k \xi_n(u).$$

ここで $u^{2k} := u_1^{2k_1} u_2^{2k_2} \dots u_r^{2k_r}$，$u^k := u_1^{k_1} u_2^{k_2} \dots u_r^{k_r}$ であり，$\xi_n(w)$ は正規確率過程 $\xi(w)$ に法則収束する．このとき正規化された損失関数は次のように表される．

$$\frac{1}{n} R_n^0(g(u)) = u^{2k} - \frac{1}{\sqrt{n}} u^k \xi_n(u) - \frac{a_n}{n} \log \varphi(g(u)).$$

次に $t \in \mathbb{R}^1$，$v = (v_1, v_2, \cdots, v_d) \in \mathbb{R}^d$ に対して $v_i = \begin{cases} \sqrt{u_i^2 - \frac{k_i}{k_a} u_a^2} & (1 \leq i \leq r) \\ u_i & (r \leq i \leq d) \end{cases}$,$t = u^{2k}$ とすると $v \in V := \{v = (v_1, v_2, \cdots, v_d) \in [0, 1]^d | v_1, v_2, \dots, v_d = 0\}$ が成り

立つ.
よって写像 $u \in [0,1]^d \to (t,v) \in T \times V$ を定める．このとき正規化された損失関数は次のように表される ([6]).

$$\frac{1}{n} R_n^0(g(t,v)) = t^2 - \frac{1}{\sqrt{n}} t\xi_n(t,v) - \frac{a_n}{n} \log \varphi(g(t,v)).$$

定義 27 (平均・経験対数損失関数)　条件付き確率 $p(x|w)$ に対して, 平均対数損失関数 $L(w)$ を次で定める.

$$L(w) := -\int q(x) \log p(x|w) dx.$$

条件付き確率 $p(x|w)$ に対して, 経験対数損失関数 $L_n(w)$ を次で定める.

$$L_n(w) := -\frac{1}{n} \sum_{i=1}^{n} \log p(x_i|w).$$

また関数近似モデル $p(y|x,w) := \frac{1}{\sqrt{2\pi}} \exp(-\frac{|y-f(x,w)|^2}{2})$ に対して, 損失関数 $L_n(w)$ を次で定める.

$$L_n(w) := -\frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i,w))^2.$$

定理 7 (汎化損失と経験損失の挙動)　対数損失関数が相対的に有限な分散をもつとする. $U_0 := \{u \in [0,1]^d | K(g(u)) = 0\}$ を満たす集合を最大にするパラメータ $\hat{u}$ を次で定める.

$$\hat{u} := \arg \max_{K(g(u))=0} \left\{ \frac{1}{4} \max\{0, \xi_n(u)\}^2 + \log \varphi(g(u)) \right\}.$$

このとき汎化損失と経験損失は次の挙動を持つ.

$$K(\hat{w}) = L(g(\hat{u})) - L(w_0) = \frac{1}{4n} \max\{0, \xi_n(\hat{u})\}^2 + o_p\left(\frac{1}{n}\right),$$

$$K_n(\hat{w}) = L_n(g(\hat{u})) - L_n(w_0) = -\frac{1}{4n} \max\{0, \xi_n(\hat{u})\}^2 + o_p\left(\frac{1}{n}\right).$$

最急降下法では t の部分は急速に最適値に近づくが v の部分はなかなか最適値に近づかない．一般には t についての最適化は汎化損失を小さくするが，v についての最適化は汎化損失を大きくする.

定義 28 (過学習)　経験損失は単調に小さくなるが，汎化損失は途中から大きくなる現象を過学習という．

予め用意した学習 (訓練) データでの正解率がいくら高くても, 実際の運用では役に立たない ([6]).

$-3 \leq x \leq 3$ の $\mathbb{R}^1$ 上の一様分布に従う確率変数 X を入力とする．平均 0, $\sigma = 0.15$ である雑音 Z に対して，$Y = 5\tanh(3x) + Z$ で定まる $\mathbb{R}^1$ 上の確率変数 Y を出力とする．ここで，出力 Y が従う条件付確率 $p(y|x,w) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{|y-f(x,w)|^2}{2\sigma^2}\right)$ を学習モデルとする．出力 Y が従う条件付確率 $q(y|x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{|y-5\tanh(3x)|^2}{2\sigma^2}\right)$ を真の分布とする．

定義 29 (2 乗誤差関数)　学習モデル $p(y|x,w)$ に対して, *2* 乗誤差関数を次で定める．

$$L_n(w) := -\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i,w))^2.$$

Mathematica を用いてテストデータとニューラルネットワークの出力の 2 乗誤差を求めるため, 次のように入力する:

$\texttt{meanTestLoss[net]} := \texttt{SquaredEuclideanDistance[net[testX], testY]/Length[testX]}$

定義 30 (RMS(二乗平均平方根) 重み)　重み w_1, $w_2, \cdots w_n$ に対して，*RMS(*二乗平均平方根*)* 重みを $\sqrt{\frac{1}{n}\sum_{i=1}^{n} {w_i}^2}$ で定める．

3.3.2　過学習を起こす場合

30 個の入力 xs，出力 ys，入力から出力への訓練データを Mathematica を用いて次のように作成する．

$\texttt{xs} = \texttt{RandomReal[UniformDistribution[}-\texttt{3, 3], 30];}$
$\texttt{ys} = \texttt{5Tanh[3xs]} + \texttt{RandomVariate[NormalDistribution[0, . 15], 30];}$
$\texttt{data} = \texttt{Normal[AssociationThread[xs}- > \texttt{ys]];}$

ここで入力を dataX = Keys[data]; 出力を dataY = Values[data]; として取り出すことができる.

30 個の入力 x, 出力 y, テストデータ $testData$ を xs, ys, $data$ と同様に作成する. ここで入力を testX = Keys[testData]; 出力を testY = Values[testData]; として取り出すことができる.

ニューラルネットワークの作成，訓練データの学習

学習モデルとして中間ユニット数を 1000 である 3 層ニューラルネットワーク（重みを 0, バイアスをなしとする.）を作成するために次のように入力すると，以下の図 3.3 の左側のように表示される.

```
net = NetChain[{LinearLayer[1000], Tanh, LinearLayer[1, "Weights" -> {Table[0, 1000]},
"Biases" -> None]}]
```

2 乗誤差を損失関数として, ニューラルネットワークに対して 50000 ラウンド学習させるために次のように入力すると，学習結果の要約が以下の図 3.3 の右側のように表示される.

```
results1 = NetTrain[net, data, All, MaxTrainingRounds -> 50000]
```

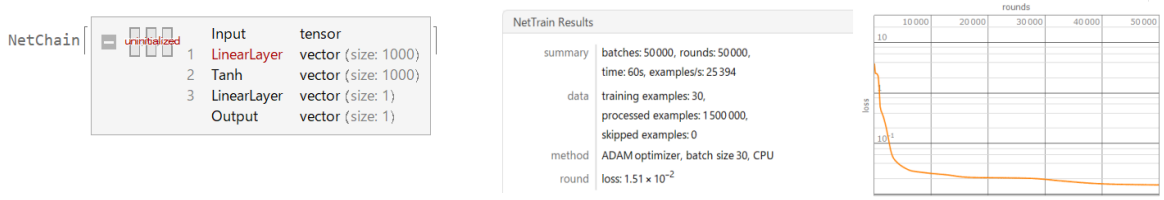


図 3.3　ニューラルネットワーク, ネットワークの学習

学習後のニューラルネットワークを overfitNet と定めて, テストデータに対する 2 乗誤差を計算するために，meanTestLoss[overfitNet] と入力すると 0.172604 と出力される．訓練データに対する経験損失 0.0151235 より大きく，過学習が起こったことが表示される.

訓練データとテストデータに対して，ニューラルネットワークの出力値を表すために次のように入力する:

```
Show[Plot[overfitNet[x], x, −3, 3], ListPlot[List@@@data, PlotStyle -> Red]]
Show[Plot[overfitNet[x], x, −3, 3], ListPlot[List@@@testData, PlotStyle -> Orange]]
```

訓練データ上への出力値は図 3.4 の左側に，テストデータ上への出力値は図 3.4 の右側に示す．学習後のニューラルネットワークはテストデータをうまく近似できない．

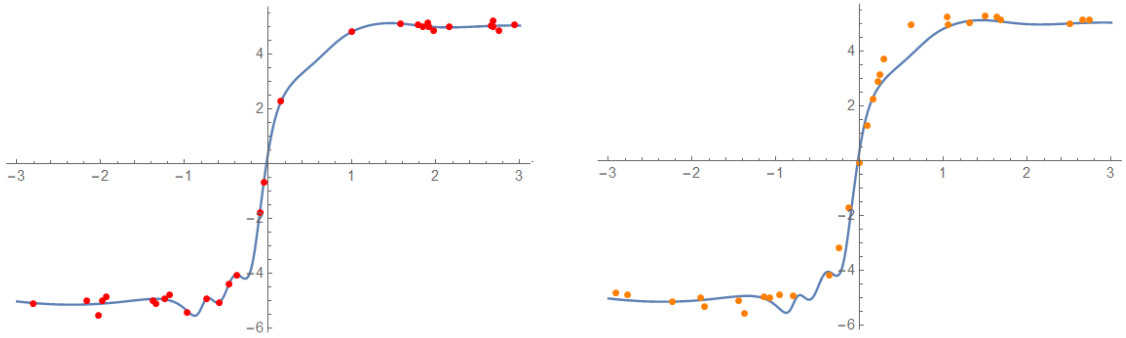


図 3.4　ニューラルネットワークの出力値 (過学習)

最大 20000 ラウンド学習する中でテストデータの損失が最小となるラウンド数で学習を終えるように，訓練データを学習させるために次のように入力する:

```
results2 = NetTrain[net, data, All, ValidationSet− > testData, MaxTrainingRounds− > 20000]
```

損失が最小であるラウンド数は 6464 であり，学習結果の要約と, 損失の変化が図 3.5 に示す．

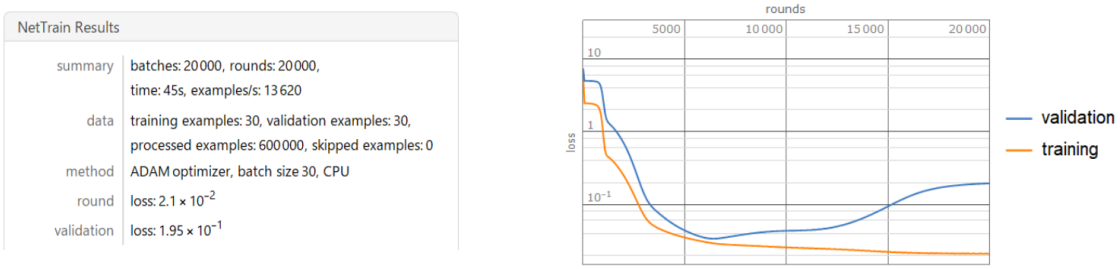


図 3.5　ニューラルネットワークの学習 (過学習)

よって早期に学習を終えたラウンドでの経験損失 0.0338247 は最終ラウンドでの経験損失 0.195436 や，テストデータに対する 2 乗誤差 0.172604 よりも小さい値である．

早期に学習を終えたニューラルネットワークを earlyStoppingNet と定めて，訓練データとテストデータに対して，ニューラルネットワークの出力値を表すために次のように入

力する:

```
Show[Plot[earlyStoppingNet[x], x, −3, 3], ListPlot[List@@@data, PlotStyle− > Red]]
Show[Plot[earlyStoppingNet[x], x, −3, 3], ListPlot[List@@@testData, PlotStyle− > Purple]]
```

訓練データ上への出力値は図 3.6 の左側に，テストデータ上への出力値は図 3.6 の右側に示す．学習後のニューラルネットワークはテストデータをうまく近似する（以下の図 3.6 の左と右の図は訓練データにもテストデータにもフィットしていることを現している．）．

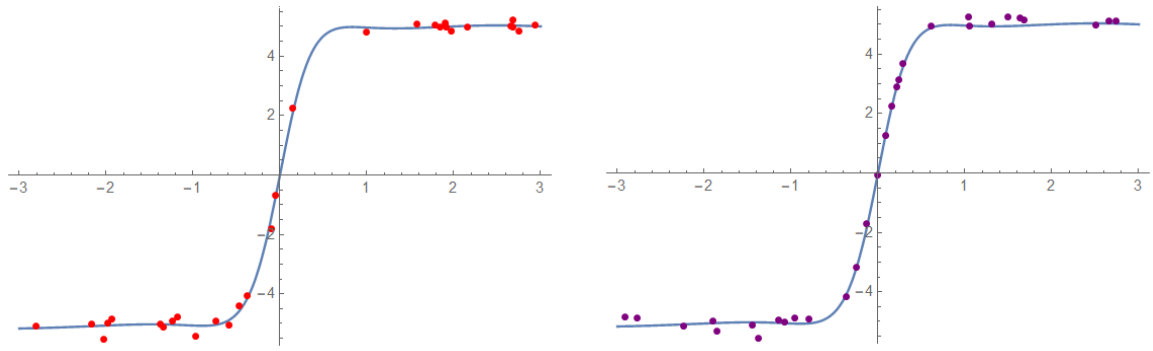


図 3.6　ニューラルネットワークの出力値のグラフ

3.3.3　過学習を起こさない場合

ニューラルネットワークの作成，訓練データの学習

学習モデルとして中間ユニット数が 2 である 3 層ニューラルネットワーク（重みを第 1 層から第 2 層へは 0 と 0，第 2 層から第 3 層へは 5 と 5，バイアスをなしとする．）を作成するために次のように入力すると，以下の図 3.7 の左側のように表示される．

```
net = NetChain[{LinearLayer[2, "Weights"− > {{5}, {5}}, "Biases"− > None],
Tanh, LinearLayer[1, "Weights"− > {{0, 0}}, "Biases"− > None]}]
```

2 乗誤差を損失関数として，ニューラルネットワークに対して 50000 ラウンド学習させるために次のように入力すると，学習結果の要約が図 3.7 の右側に表示される．

```
results1 = NetTrain[net, data, All, MaxTrainingRounds− > 50000]
```

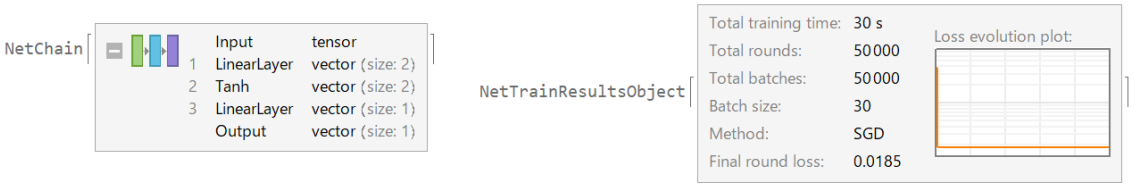


図 3.7　ニューラルネットワーク, ネットワークの学習

学習後のニューラルネットワークを fitNet と定めて, テストデータに対する 2 乗誤差を計算するために，`meanTestLoss[fitNet]` と入力すると 0.0202818 と出力される．訓練データに対する経験損失 0.0184928 と比べて余り大きくなく, 過学習が起こらなかったことが表示される．

訓練データとテストデータに対して，ニューラルネットワークの出力値を表すために次のように入力する:

```
Show[Plot[fitNet[x], x, −3, 3], ListPlot[List@@@data, PlotStyle− > Red]]
Show[Plot[fitNet[x], x, −3, 3], ListPlot[List@@@testData, PlotStyle− > Orange]]
```

訓練データ上への出力値は図 3.8 の左側に，テストデータ上への出力値は図 3.8 の右側に示す．学習後のニューラルネットワークはテストデータを近似する．

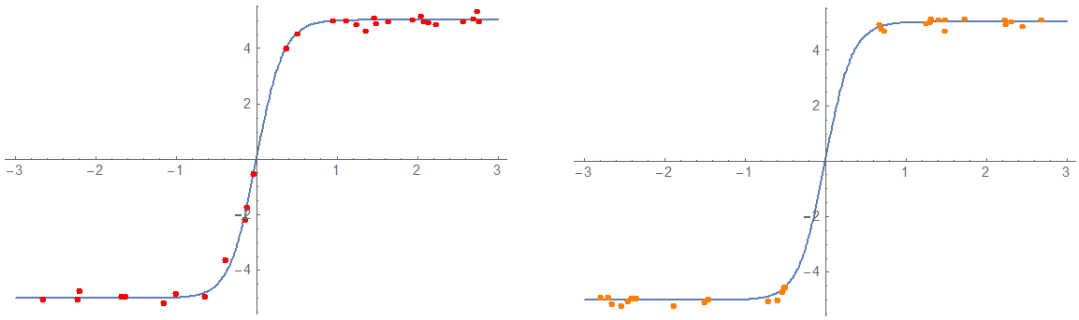


図 3.8　ニューラルネットワークの出力値

一般に，中間ユニット数が大きいと過学習が起こることが知られている ([6])．過学習を起こす場合 (3.3.2 参照) として中間ユニット数を 1000 として実行したが，現部分節 3.3.3 では過学習を起こさないように中間ユニット数を 2 として Mathematica を用いて実行した．

RMS(二乗平均平方根) 重みのグラフ

10000 ラウンド学習する過程で変化する RMS 重みをグラフに表すために次のように入力すると，以下の図 3.9 の左側に表示される．

```
results2 = NetTrain[net, data, "RMSWeightsEvolutionPlot", MaxTrainingRounds− > 10000]
```

2000 ラウンド学習する過程で変化する RMS 重みの値を，次のように入力して求める．

```
results3 := NetTrain[net, data, "RMSWeightsHistories", MaxTrainingRounds− > 2000]
```

学習前のニューラルネットワークを net1 と定めて, RMS 重みの初期値を，次のように入力して求める．

```
N1 = RootMeanSquare[Flatten[NetExtract[net1, 1, "Weights"]]];
N2 = RootMeanSquare[Flatten[NetExtract[net1, 3, "Weights"]]];
```

スケールを変更せずに RMS 重みの変化をグラフに表すために次のように入力すると，以下の図 3.9 の右側に表示される．

```
t1 = TimeSeries[Join[N1, results3[1, "Weights"]]];
t2 = TimeSeries[Join[N2, results3[3, "Weights"]]];
ListLinePlot[t1, t2, PlotRange− > All, PlotLabels− > 1, "Weights", 3, "Weights"]
```

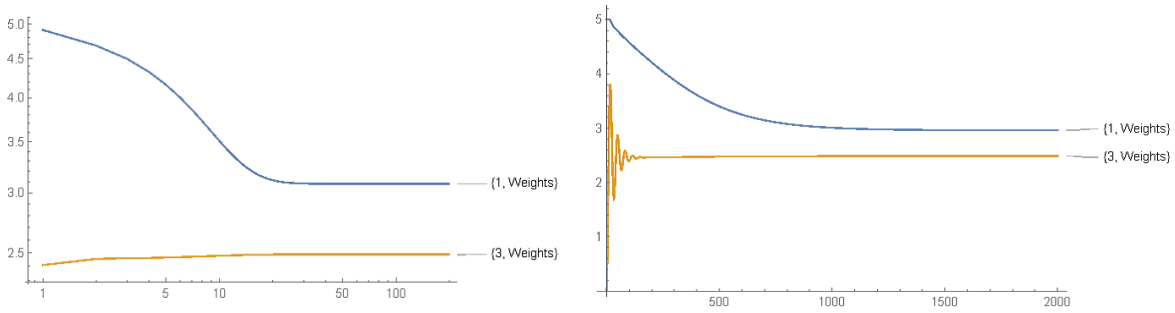


図 3.9　RMS 重みのグラフ

重みの収束値

10000 ラウンド学習させて最終ラウンドでの重みの値を求めるために次のように入力する:

```
result4 = NetTrain[net, data, "FinalWeights", MaxTrainingRounds− > 10000]
```

以下のように出力され，第 1 層から第 2 層への重みは約 3 であることが表示される．

```
< |{1, "Weights"}− > {{2.98904}, {2.98904}}, {3, "Weights"}− > {{2.49893, 2.49893}}| >
```

このとき第 2 層から第 3 層への重みの和を求めるために次のように入力する:

```
w3 := Flatten[results4[3, "Weights"]]
RealSign[First[results4[1, "Weights"]]] * First[w3] +
RealSign[Last[results4[1, "Weights"]]] * Last[w3]
```

4.99974 と出力され，第 2 層から第 3 層への重みの和は約 5 であることが表示される.

第 4 章

パラメータの表示の一般化

真の分布を実現するパラメータを一般化させることでパラメータ空間に現れる特異点の複雑さが変わる．学習が進むとプラトーが起こるが, 特異点の複雑さが下がり (構造発見) 学習が進む．初めに，真の分布を実現するパラメータ表示を一般化して求める．次に，特異点解消定理を用いて汎化誤差を求める手法について示し，簡単な場合 (中間ユニット数 $H = 2 \to H_0 = 0$ や $H = 2 \to H_0 = 1$ の場合) の特異点やカルバック情報量や学習係数を求める．

7 章以降で中間ユニット数 $H = 2 \to H_0 = 1$ の場合について，数学教育の過剰般化現象に応用させて考察する（図 4.1）．複雑な特異点現象に対する数学教育への応用については今後の課題として, 第 10 章に後述する．

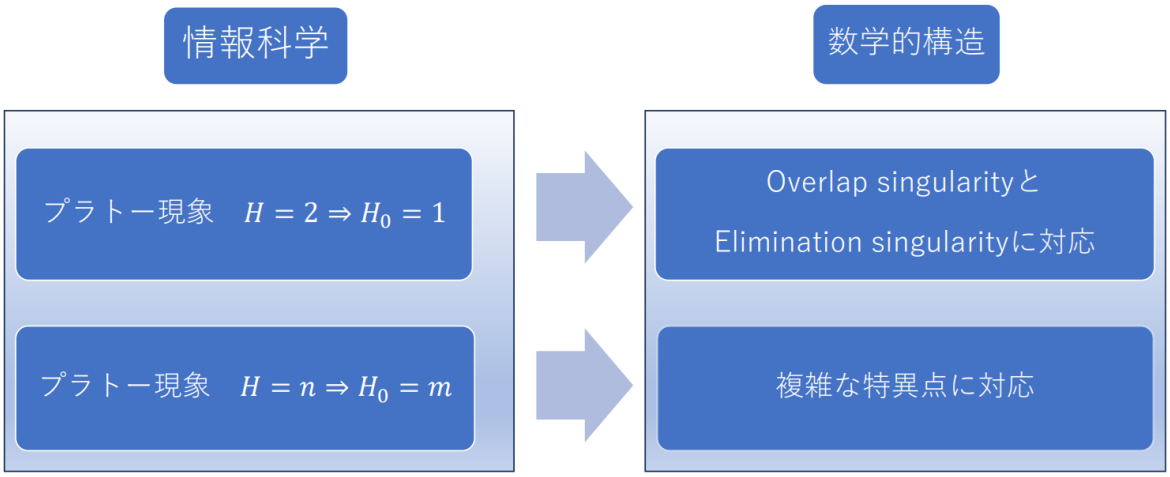


図 4.1　本研究におけるパラメータの一般化の位置づけ

4.1 ヒルベルトの基底定理

4.1.1 代数的集合とイデアル

本研究では以下の定義 31 から定義 36，定理 8,9 を用いて代数的な考察を行う（[22]）.

定義 31 (イデアル) 部分集合 $I \subset \mathcal{R} = \mathbb{R}[x_1, x_2, \cdots, x_d]$ がイデアルであるとは次の条件を満たすときと定める.

$$f(x), g(x) \in I \Longrightarrow f(x) + g(x) \in I,$$
$$f(x) \in I, g(x) \in \mathcal{R} \Longrightarrow f(x) g(x) \in I.$$

定義 32 (有限生成イデアル) 多項式の集合 $f_1, f_2, \cdots, f_n$ に対して，次で定まる $\mathcal{R}$ の部分集合

$$\langle f_1, f_2, \cdots, f_n \rangle = \left\{ \sum_{i=1}^{k} g_i(x) f_i(x) ; g_i \in \mathcal{R} \right\},$$

は $f_1, f_2, \cdots, f_n$ を含む最小のイデアルとし，このイデアルを

$$f_1, f_2, \cdots, f_n$$

で生成されるイデアルと定める.

定義 33 (解析的集合) U を $\mathbb{R}^d$ の開集合，$f: U \longrightarrow \mathbb{R}$ を解析関数とする．f の零点全体のつくる集合を解析的集合と定める.

$$\{x \in U | f(x) = 0\}.$$

解析関数 $f_1(x), f_2(x), \cdots, f_k(x)$ がすべて零になる点全体の集合も解析的集合と定める.

$$\{x \in U | f_1(x) = f_2(x) = \cdots = f_k(x) = 0\}.$$

定義 34 (代数的集合)　イデアル $I \subset \mathcal{R} = \mathbb{R}[x_1, x_2, \cdots, x_d]$ に対して，代数的集合 $V(I)$ を次で定める．

$$V(I) = \left\{ x \in \mathbb{R}^d \middle| f(x) = 0 \, (\forall f \in I) \right\}.$$

$V = V(I)$ を満たすイデアル I が存在するとき $V(I) \subset \mathbb{R}^d$ を代数的集合と呼ぶ.

定義 35 (代数的集合 V によって定まるイデアル)　代数的集合 $V \subset \mathbb{R}^d$ に対して，全ての多項式が零になる V 上の集合 $I(V)$ を次で定める．

$$I(\mathrm{V}) = \{ f(x) \in \mathbb{R}[x_1, x_2, \cdots, x_d] | f(x) = 0 \, (\forall x \in \mathrm{V}) \}.$$

$I(V) \subset \mathbb{R}[x_1, x_2, \cdots, x_d]$ はイデアルの定義を満たす．$I(V)$ を代数的集合 V によって定まるイデアルという．

定義 36 (根基イデアル)　イデアル $I \subset \mathbb{R}[x_1, x_2, \cdots, x_d]$ に対して，集合

$$\sqrt{I} = \{ f(x) \in \mathbb{R}[x_1, x_2, \cdots, x_d] | f(x)^m \in I \text{ となる自然数 } m \text{ が存在する．} \}$$

はイデアルになり，このイデアルを I の根基という．$I = \sqrt{I}$ が成り立つときイデアル I は根基イデアルという．

定理 8 (ヒルベルトの零点定理)　代数閉体 k，イデアル $I \subset \mathbb{R}[x_1, x_2, \cdots, x_d]$ に対して，次が成り立つ.

$$I(V(I)) = \sqrt{I}.$$

定理 9 (イデアルと代数的集合の対応)　代数閉体 k に対して，2 つの写像 V : 代数的集合 $\to$ 根基イデアル と I : 根基イデアル $\to$ 代数的集合 は全単射であり，互いに逆写像である．

4.1.2 グレブナ基底と消去イデアル

本研究では以下の定義 37 から定義 40, 定理 10 から定理 12 を用いて計算代数としての考察を行う（[23]）.

定義 37 (単項式順序) $k[x_1, x_2, \cdots, x_d]$ における単項式順序とは, $\mathbb{Z}^n_{\geq 0}$ の順序付け $>$, あるいは単項式の集合 $\{x^\alpha | \alpha \in \mathbb{Z}^n_{\geq 0}\}$ の順序付けで, 次の性質を満たすものである.
(i) $>$ は $\mathbb{Z}^n_{\geq 0}$ の全順序である.
(ii) $\alpha > \beta$ で $\gamma \in \mathbb{Z}^n_{\geq 0}$ とすれば, $\alpha + \gamma > \beta + \gamma$ である.
(iii) $>$ は $\mathbb{Z}^n_{\geq 0}$ の整列順序である. これは $\mathbb{Z}^n_{\geq 0}$ のどんな空でない部分集合も, $>$ に関する最小元を持つということである.

定義 38 (先頭項) 体 k に対し, $f \in k[x_1, x_2, \cdots, x_d]$ をゼロでない多項式とする. f に現れる単項式のなかで単項式順序 $>$ に関して最大のものを先頭項とよび, $LT(f)$ と表す.

定義 39 (グレブナ基底) 単項式順序を固定する. イデアル I の有限部分集合 $G = \{g_1, \cdots, g_t\}$ がグレブナ基底であるとは,

$$\langle LT(g_1), \cdots, LT(g_t) \rangle = \langle LT(I) \rangle$$

を満たすことと定める.

定義 40 (消去イデアル) イデアル $I = \langle f_1, \cdots, f_s \rangle \subset k[x_1, x_2, \cdots, x_d]$ に対して, l 次の消去イデアル I_l であるとは,

$$I_l := I \cap k[x_{l+1}, \cdots, x_d],$$

で定まる $k[x_{l+1}, \cdots, x_d]$ のイデアルである.

定理 10 (消去定理) $I \subset k[x_1, x_2, \cdots, x_d]$ をイデアルとし, G を I の単項式順序 $>$ に関するグレブナ基底とする.

$0 \leq l \leq n$ に対して，集合

$$G_l := G \cap k[x_{l+1}, \cdots, x_d],$$

は l 次の消去イデアル I_l のグレブナ基底である．

定理 11 (昇鎖条件)

$$I_1 \subset I_2 \subset I_3 \subset \cdots$$

を $k[x_1, x_2, \cdots, x_d]$ のイデアルの昇鎖とする．このとき，整数 $N \geq 1$ が存在して，

$$I_N = I_{N+1} = I_{N+2} = \cdots .$$

となる．

定理 12 (ヒルベルトの基底定理) イデアル $I \subset k[x_1, x_2, \cdots, x_d]$ は有限個の生成集合をもつ．ある $g_1, \cdots, g_t \in I$ に対して $I = \langle g_1, \cdots, g_t \rangle$ となる．

4.2 基底定理を応用した補題

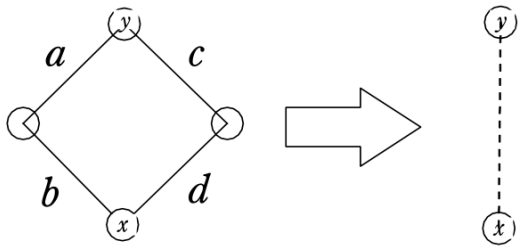
以下の結果は [24]，副論文 [25] に基づく．

4.2.1 中間ユニット数 $H = 2 \to H_0 = 0$ の場合

平均 0, $\sigma = 1$ である雑音 Z に対して，学習モデルを入力ユニット数 1，中間ユニット数 $H = 2$，出力ユニット数 1，活性化関数 $\tanh(x)$ である 3 層パーセプトロンとし，真の分布を中間ユニット数 $H_0 = 0$ である 3 層パーセプトロンとする ([24]).
$f(x, w)$ と真の分布を次で定める．

$$f(x, w) = a \tanh(bx) + c \tanh(dx),\ q(x, y) = \frac{q(x)}{\sqrt{2\pi}} \exp(-\frac{|y|^2}{2}).$$

図 4.2 は真の分布が $H_0 = 0$ で実現される場合を表す．



真の分布が $H_0=0$ で実現される場合

図 4.2　$H = 2 \Rightarrow H_0 = 0$ の場合

学習モデルと真の分布が等しくなるパラメータの解析的集合を考える.

$$W_0 := \{w \in \mathbb{R}^4 | p(x, y|w) = q(x, y)\} = \{w \in \mathbb{R}^4 | a\tanh(bx) + c\tanh(dx) = 0\}.$$

$\tanh(x)$ を級数展開すると，定義方程式は次で表される.

$$\sum_{k=0}^{\infty} \frac{2^{2k+2}(2^{2k+2}-1)B_{2k+2}x^{2k+1}}{(2k+2)!}(ab^{2k+1} + cd^{2k+1}).$$

ここで，B_n はベルヌーイ数である．$\{x^{2k+1}\}$ は一次独立であるから，W_0 は $g_k(a, b, c, d) := ab^{2k+1} + cd^{2k+1}$ と定めると, 無限個の多項式の共通零点である.

$$W_0 = \{w \in \mathbb{R}^4 | g_0 = g_1 = \cdots = 0\}.$$

イデアル $I_k = \langle g_0, g_1, g_2, \cdots, g_k \rangle$ とすると，イデアルの増大列 $I_0 \subset I_1 \subset I_2 \subset I_3 \subset \cdots$ は止まる.

補題 1 ([24])　4 変数多項式 $F_k(a_1, b_1, a_2, b_2) := a_1{b_1}^{2k+1} + a_2{b_2}^{2k+1}$ に対して, イデアル $I_k = \langle F_0, F_1, F_2, \cdots, F_k \rangle$ とする．このときイデアルの増大列 $I_0 \subset I_1 \subset I_2 \subset I_3 \subset \cdots$ は止まり, 任意の $k \geq 1$ に対して, $I_1 = I_k$ が成り立つ.

証明　*Mathematica* を用いて，F_2, $F_3 \in \langle F_0, F_1 \rangle$ が成り立つ.
また, 次の漸化式が成り立つ.

$$F_{k+1} = F_1(b_1^{2k} + b_2^{2k}) - F_0(b_1^2 b_2^{2k} + b_1^{2k} b_2^2) + (b_1^2 b_2^2)F_{k-1}.$$

よって任意の $k \geq 1$ に対して, $F_k \in \langle F_0, F_1 \rangle$ が成り立つ.　□

補題 1 より次が成り立つ.

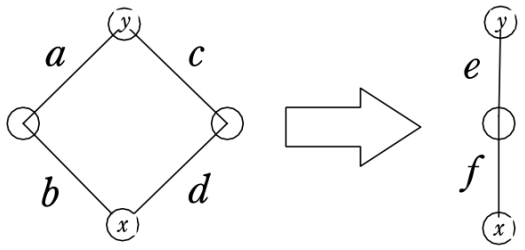
$$V(I_k) = \{w \in \mathbb{R}^4 | g_0 = g_1 = \cdots = g_k = 0\} = \{w \in \mathbb{R}^4 | ab + cd = ab^3 + cd^3 = 0\}.$$

4.2.2 中間ユニット数 $H = 2 \to H_0 = 1$ の場合

平均 0, $\sigma = 1$ である雑音 Z に対して，学習モデルを入力ユニット数 1，中間ユニット数 $H = 2$，出力ユニット数 1，活性化関数 $\tanh(x)$ である 3 層パーセプトロンとし，真の分布を中間ユニット数 $H_0 = 1$ である 3 層パーセプトロンとする [24].
$f(x, w)$ と真の分布を次で定める.

$$f(x, w) = a\tanh(bx) + c\tanh(dx),\ q(x, y) = \frac{q(x)}{\sqrt{2\pi}} \exp(-\frac{|y - e\tanh(fx)|^2}{2}).$$

図 4.3 は真の分布が $H_0 = 1$ で実現される場合を表す.



真の分布が $H_0 = 1$ で実現される場合

図 4.3　$H = 2 \Rightarrow H_0 = 1$ の場合

学習モデルと真の分布が等しくなるパラメータの解析的集合を考える.

$$W_0 := \{w \in \mathbb{R}^4 | p(x, y|w) = q(x, y)\} = \{w \in \mathbb{R}^4 | a\tanh(bx) + c\tanh(dx) = e\tanh(fx)\}.$$

$\tanh(x)$ を級数展開して，$g_k(a, b, c, d, e, f) := ab^{2k+1} + cd^{2k+1} - ef^{2k+1}$ と定めると，定義方程式は次で表される.

$$\sum_{k=0}^{\infty} \frac{2^{2k+2}(2^{2k+2} - 1)B_{2k+2}x^{2k+1}}{(2k + 2)!} g_k(a, b, c, d, e, f).$$

ここで B_n はベルヌーイ数である．$\{x^{2k+1}\}$ は一次独立であるから, W_0 は無限個の多項式の共通零点である．

$$W_0 := \{w \in \mathbb{R}^4 | g_0 = g_1 = \cdots = 0\}.$$

イデアル $I_k = \langle g_0, g_1, g_2, \cdots, g_k \rangle$ とすると，イデアルの増大列 $I_0 \subset I_1 \subset I_2 \subset I_3 \subset \cdots$ は止まる．

補題 2 ([24]) 6 変数多項式 $F_k(a_1, a_2, a_3, b_1, b_2, b_3) := a_1 {b_1}^{2k+1} + a_2 {b_2}^{2k+1} + a_3 {b_3}^{2k+1}$ に対して，イデアル $I_k = \langle F_0, F_1, F_2, \cdots, F_k \rangle$ とする．このときイデアルの増大列 $I_0 \subset I_1 \subset I_2 \subset I_3 \subset \cdots$ は止まり, 任意の $k \geq 2$ に対して, $I_2 = I_k$ が成り立つ.

証明 *Mathematica* を用いて,F_3, F_4, $F_5 \in \langle F_0, F_1, F_2 \rangle$ が成り立つ.
次の漸化式が成り立つ.

$$\begin{aligned} F_{k+2} = {} & F_2({b_1}^{2k} + {b_2}^{2k} + {b_3}^{2k}) - F_1({b_1}^2 d^{2k} + {b_1}^2 {b_3}^{2k} + {b_1}^{2k} {b_2}^2 \\ & + {b_2}^2 {b_3}^{2k} + {b_1}^{2k} {b_3}^2 + {b_2}^{2k} {b_3}^2) + F_0({b_1}^2 {b_2}^2 {b_3}^{2k} + {b_1}^2 {b_2}^{2k} {b_3}^2 + {b_1}^{2k} {b_2}^2 {b_3}^2) \\ & + ({b_2}^2 {b_3}^2 + {b_1}^2 {b_3}^2 + {b_1}^2 {b_2}^2) F_k - 2({b_1}^2 {b_2}^2 {b_3}^2) F_{k-1}. \end{aligned}$$

$k \geq 2$ に対して，$F_k \in \langle F_0, F_1, F_2 \rangle$ が成り立つ. □

補題 2 より次が成り立つ.

$$\begin{aligned} & V(I_k) = \{w \in \mathbb{R}^4 | g_0 = g_1 = \cdots = g_k = 0\} \\ & = \{w \in \mathbb{R}^4 | ab + cd - ef = ab^3 + cd^3 - ef^3 = ab^5 + cd^5 - ef^5 = 0\}. \end{aligned}$$

4.2.3 解析的集合の定義方程式の有限性

拡張された補題の証明を以下で示す.

定義 41 (副論文 [25]) $2d$ 変数 $(a_1, \cdots, a_d, b_1, \cdots, b_d)$ 多項式 $F_n(a_1, \cdots, a_d, b_1, \cdots, b_d) := \sum_{k=1}^{d} a_k \, b_k^{2n+1}$ と定める.

補題 3 (副論文 [25]) $2d$ 変数多項式 $F_n(a_1, \cdots, a_d, b_1, \cdots, b_d)$ について，イデアル $I_k = \langle F_0, F_1, F_2, \cdots, F_k \rangle$ とする．このときイデアルの増大列 $I_0 \subset I_1 \subset I_2 \subset I_3 \subset \cdots$ は止まり，任意の $k \geq d-1$ に対して，$I_{d-1} = I_k$ が成り立つ．

4.3 真の分布を実現するパラメータ表示

以下の結果は副論文 [26] に基づく．

定義 42 ([6]) $W \subset \mathbb{R}^d$ をパラメータ全体の集合とする．あるパラメータ $w \in W$ が存在して，$q(x) = p(x|w)$ を満たすとき，$q(x)$ は $p(x|w)$ により実現可能であるという．そうでないとき実現可能でないという．真のパラメータの集合 W_{00} を次で定める．

$$W_{00} = \{w \in W | \text{ すべての } x \text{ について } q(x) = p(x|w)\}.$$

補題 4 ([6])
(1) 真の分布 $q(x)$ が学習モデル $p(x|w)$ によって実現可能であることと，真のパラメータの集合 W_{00} が空集合でないことは同値である．
(2) W_{00} が空集合でないとする．W_{00} の要素は一つとは限らないが，任意の $w \in W_{00}$ について，$p(x|w)$ は同じ確率分布を表している．

平均対数損失 $L(w)$ が小さい値であればあるほど $p(x|w)$ は $q(x)$ をよく近似していると考えてよい．特に真のパラメータの集合はカルバック情報量が 0 になるパラメータの集合であり，次が成り立つ ([6]).

$$W_{00} = \left\{ w \in W | \int q(x) \log \frac{q(x)}{p(x|w)} dx = 0 \right\}.$$

定義 43 ([6]) パラメータの集合を $W \subset \mathbb{R}^d$ とし，平均対数損失関数 $L(w)$ を最小にするパラメータの集合を W_0 とする．

$$W_0 = \{w \in W | L(w) \text{ が最小値をとる } \}.$$

この集合を真の分布に対して最適なパラメータ集合と呼ぶ．集合 W_0 の要素が w_0 が *1* つのみで，w_0 を含む開集合で W に含まれるものが存在して，w_0 でのヘッセ行列 $\nabla^2 L(w_0)$

すなわち，$d \times d$ 行列でその ij 成分が

$$\left\{\nabla^2 L(w_0)\right\}_{ij} = \left\{\frac{\partial^2 L}{\partial w_i \partial w_j}\right\}(w_0),$$

で定義される行列が正則 (固有値が全て正の値である) であるとき，$q(x)$ は $p(x|w)$ に対して正則であるという．正則でないとき $q(x)$ は $p(x|w)$ に対して正則でないという．

真の分布 $q(x)$ が学習モデル $p(x|w)$ によって実現可能であるとき次が成り立つ．

$$W_{00} = W_0.$$

真の分布 $q(x)$ が学習モデル $p(x|w)$ によって実現可能でないとき次が成り立つ．

$$W_{00} \neq W_0.$$

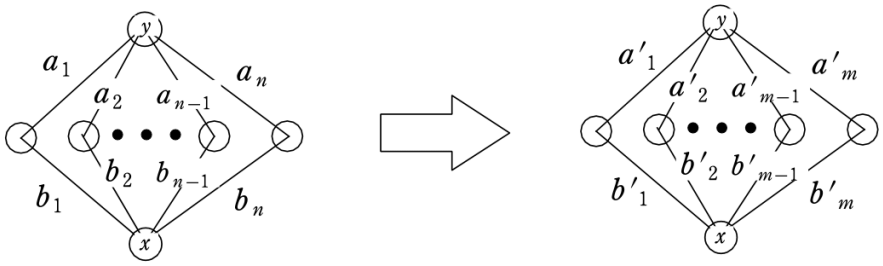
4.3.1 中間ユニット数 $H = n \to H_0 = m$ の場合

平均 0, $\sigma = 1$ である雑音 Z に対して，学習モデルを入力ユニット数 1，中間ユニット数 $H = n$，出力ユニット数 1，活性化関数 $\tanh(x)$ である 3 層パーセプトロンとし，真の分布を中間ユニット数 $H_0 = m$ である 3 層パーセプトロンとする．
$f(x, w)$ と真の分布を次で定める ([24]).

$$f(x, w) = a_1 \tanh(b_1 x) + \cdots + a_n \tanh(b_n x)$$

$$q(x, y) = \frac{q(x)}{\sqrt{2\pi}} \exp(-\frac{|y - a'_1 \tanh(b'_1 x) + \cdots + a'_m \tanh(b'_m x)|^2}{2}).$$

図 4.4 は真の分布が $H_0 = m$ で実現される場合を表す．



真の分布が $H_0 = m$ で実現される場合

図 4.4　$H = n \Rightarrow H_0 = m$ の場合

真の分布が学習モデルによって実現されるパラメータの集合を考える

$$W_0 := \{\ w\ \in \mathbb{R}^{2n} \mid p(x,y|w) = q(x,y)\}.$$

学習モデルを 1 つの入力層，中間ユニット $H = n$，1 つの出力ユニットである 3 層ニューラルネットワーク活性化関数を tanh とし，真の分布を中間ユニット $H = m$ である 3 層ニューラルネットワークとする．

$$\begin{aligned} W_0 &:= \{\ w\ \in \mathbb{R}^{2n} \mid p(x,y|w) = q(x,y)\} \\ &= \{\ w\ \in \mathbb{R}^{2n} \mid a_1 \tanh(b_1 x) + \cdots + a_n \tanh(b_n x)\ = a'_1 \tanh\left(b'_1 x\right) + \cdots + a'_m \tanh\left(b'_m x\right)\}. \end{aligned}$$

$\tanh(x)$ は奇関数より b_i は条件 $b_i \geq 0$ を満たす．

活性化関数のテイラー展開をして次を得る．

$$\tanh(x) := \sum_{k=1}^{\infty} \frac{2^{2k}\left(2^{2k}-1\right) B_{2k} x^{2k-1}}{(2k)!} = x - \frac{x^3}{3} + \frac{2x^5}{15} - \frac{17x^7}{315} + \cdots,$$

ここでベルヌーイ数 B_n を級数展開して次で定める．

$$B_0 = 1, B_n = -\frac{1}{n+1} \sum_{k=1}^{n-1} \binom{n+1}{k}.$$

解析的集合の定義方程式を考える．

$$a_1 \tanh(b_1 x) + \cdots + a_n \tanh(b_n x) - a'_1 \tanh(b'_1 x) - \cdots - a'_m \tanh(b'_m x).$$

テイラー展開をすることにより次で表される．

$$a_1 \sum_{k=1}^{\infty} \frac{2^{2k}\left(2^{2k}-1\right) B_{2k} x^{2k-1}}{(2k)!} b_1^{2k-1} + \cdots + a_n \sum_{k=1}^{\infty} \frac{2^{2k}\left(2^{2k}-1\right) B_{2k} x^{2k-1}}{(2k)!} b_n^{2k-1}$$

$$- a'_1 \sum_{k=1}^{\infty} \frac{2^{2k}\left(2^{2k}-1\right)B_{2k}x^{2k-1}}{(2k)!} {b'}_1^{2k+1} - \cdots - a'_m \sum_{k=1}^{\infty} \frac{2^{2k}\left(2^{2k}-1\right)B_{2k}x^{2k-1}}{(2k)!} {b'}_m^{2k+1}$$

$$\sum_{k=1}^{\infty} \frac{2^{2k}\left(2^{2k}-1\right)B_{2k}x^{2k-1}}{(2k)!}(a_1 b_1^{2k-1} + \cdots + a_n b_n^{2k-1})$$

$$- \sum_{k=1}^{\infty} \frac{2^{2k}\left(2^{2k}-1\right)B_{2k}x^{2k-1}}{(2k)!}(a'_1 {b'}_1^{2k+1} + \cdots + a'_m {b'}_m^{2k+1})$$

$$= \sum_{k=1}^{\infty} \frac{2^{2k}\left(2^{2k}-1\right)B_{2k}x^{2k-1}}{(2k)!} \quad (a_1 b_1^{2k-1} + \cdots + a_n b_n^{2k-1} - a'_1 {b'}_1^{2k+1} - \cdots - a'_m {b'}_m^{2k+1})$$

次の方程式を得る

$$\sum_{k=1}^{\infty} \frac{2^{2k}\left(2^{2k}-1\right)B_{2k}x^{2k-1}}{(2k)!}(\sum_{i=1}^{n} a_i b_i^{2k+1} - \sum_{i=1}^{m} {a'}_i {b'}_i^{2k+1}).$$

$a_1, \cdots, a_n, b_1, \cdots, b, a'_1, \cdots, a'_m,$ の関数 g_k は次で定まる.

$$g_k(a_1, \cdots, a_n, b_1, \cdots, b_n, a'_1, \cdots, a'_m, b'_1, \cdots, b'_m) := \sum_{i=1}^{n} a_i b_i^{2k+1} - \sum_{i=1}^{m} {a'}_i {b'}_i^{2k+1}$$

g_k 用いて方程式を次で表す.

$$\sum_{k=0}^{\infty} \frac{2^{2k+2}\left(2^{2k+2}-1\right)B_{2k+2}x^{2k+1}}{(2k+2)!} g_k = 0,$$

ここで g_k は $a_1, \cdots, a_n, b_1, \cdots, b, a'_1, \cdots, a'_m,$ の関数である.
集合 $\{x^{2k+1}\}$ は線形独立より，W_0 は有限個の多項式によって定まる共通零点である.

$$W_0 := \{\ w\ \in \mathbb{R}^{2n} \mid g_0 = g_1 = \cdots = 0\}.$$

イデアル I_k を次で定める

$$I_k = \langle g_0, g_1, g_2, \cdots, g_k \rangle .$$

このとき，非降鎖イデアルの列を定める.

4.3.2 I_{d-1} の消去イデアルのグレブナ基底

補題 5 を消去イデアルの定理を用いて示す．Mathematica を用いてグレブナ基底の消去イデアルの基底を計算した.

補題 5 (副論文 [26]) (定義方程式) $n \geq 2$ に対して, イデアル I_{d-1} を次で定める.

$$\sum_{i=1}^{n} a_i b_i = \sum_{i=1}^{n} a_i b_i^3 = \cdots = \sum_{i=1}^{n} a_i b_i^{2d-1} = 0.$$

$b_j (1 \leq j \leq n, except i)$ に対して，I_{d-1}の$a_j (1 \leq j \leq n,\ except\ i)$ の消去イデアルの定義方程式は次で定まる.

$$a_i b_i \left(b_i^2 - b_1^2\right)\left(b_i^2 - b_2^2\right) \cdots \left(b_i^2 - b_n^2\right) = 0.$$

証明　方程式を数学的帰納法によって証明する．$n = 2$ のとき

$$a_1 b_1 + a_2 b_2 = a_1 b_1^3 + a_2 b_2^3 = 0.$$

Mathematica を用いてグレブナ基底の消去イデアルの基底を計算するために次のように入力する*:*

$$\mathtt{f1} = \mathtt{a_1 b_1} + \mathtt{a_2 b_2}; \mathtt{f2} = \mathtt{a_1 b_1^3} + \mathtt{a_2 b_2^3}.$$

そのときグレブナ基底の a_2 の消去イデアルを計算するために次のように入力する*:*

$$\mathtt{GroebnerBasis}[\{\mathtt{f1},\ \mathtt{f2}\}, \{\mathtt{a_1}, \mathtt{b_1}, \mathtt{a_2}, \mathtt{b_2}\},\ \{\mathtt{a_2}\}$$

$$\mathtt{MonomialOrder} \rightarrow \mathtt{exicographic}].$$

出力は

$$\mathtt{a_1 b_1^3} - \mathtt{a_1 b_1 b_2^2},$$

次のように因数分解される.

$$a_1 b_1^3 - a_1 b_1 b_2^2 = a_1 b_1 \left(b_1^2 - b_2^2\right) = a_1 b_1 \left(b_1^2 - b_2^2\right)\ = a_1 b_1 \left(b_1 + b_2\right)\left(b_1 - b_2\right) = 0.$$

$1 \leq i\ \leq d$ に対して, $n = k$ の場合を仮定する.

次が成り立つ.

$$a_1 b_1^{2i-1} + \cdots + a_k b_k^{2i-1} = 0$$

$1 \leq i \ \leq d$ に対して, $n = k+1$ の場合を仮定する. 次が成り立つ.

$$a_1 b_1^{2i-1} + \cdots + a_{k+1} b_{k+1}^{2i-1} = 0.$$

a_{k+1} を消去することで次を得る.

$$a_1 b_1^{2i-1} + \cdots + a_{k-1} b_{k-1}^{2i-1}$$
$$- \left(a_1 b_1^{2i-3} + \cdots + a_{k-1} b_{k-1}^{2i-3}\right) b_{k+1}^2 = 0.$$

次の方程式を得る.

$$a_1 \left(b_1^2 - b_{k+1}^2\right) b_1^{2i-3} + a_2 \left(b_2^2 - b_{k+1}^2\right) b_2^{2i-3} + \cdots + a_k \left(b_k^2 - b_{k+1}^2\right) b_k^{2i-3} = 0.$$

ここで $a_j \left(b_j^2 - b_{k+1}^2\right) = c_j$ と置き換えをすると

$$c_1 b_1^{2i-3} + \cdots + c_k b_k^{2i-3} = 0.$$

この置き換えを用いて $n = k$ の場合次を得る.

$$c_i b_i \left(b_i^2 - b_1^2\right)\left(b_i^2 - b_2^2\right) \cdots \left(b_i^2 - b_k^2\right) = 0.$$

$n = k+1$ の場合次を得る.

$$a_i b_i \left(b_i^2 - b_1^2\right)\left(b_i^2 - b_2^2\right) \cdots \left(b_i^2 - b_k^2\right)\left(b_i^2 - b_{k+1}^2\right) = 0.$$

□

4.3.3 記号の準備

$w \in \mathbb{R}^{2n}$ を次で定める.

$$w := \{a_1, b_1, a_2, b_2, \ldots, a_n, b_n\}.$$

初めに l_k を $b_i = b'_k \, (1 \leq k \leq m)$ を満たす数 $b_i \, (1 \leq i \leq n)$ とする.

次を定める.

$$0 := n'_0, \sum_{k=1}^{i} l_k := n'_i.$$

次の条件を満たすように (a_i, b_i) を並びかえる.

$$b_i = b'_k \left(n'_{i-1} + 1 \leq i \leq n'_i\right).$$

次を定める.

$$\sum_{i=1}^{m} l_i := n'.$$

次に z を $a_i b_i = 0$ を満たす数 $b_i \, (n' + 1 \leq i \leq n)$ とする.
また $n' + z := n''$ を定める. 次の条件を満たすように (a_i, b_i) を並びかえる.

$$a_i b_i = 0 \, (n' + 1 \leq i \leq n''), a_i b_i \neq 0 \, (n'' + 1 \leq i \leq n).$$

最後に次に r_k を $b_i = b_k \, (n'' + 1 \leq k \leq n'' + h)$ を満たす数 $b_i \, (n'' + 1 \leq i \leq n)$ とする. $r_1 \geq r_2 \geq \cdots \geq r_h \geq 2$ を満たす. 次を定める.

$$n'' := n''_0, n'' + \sum_{k=1}^{i} (r_k + 1) := n''_i.$$

次の条件を満たすように (a_i, b_i) を並びかえる.

$$b_i = b_k \left(n''_{i-1} + 1 \leq i \leq n''_i\right).$$

次が成り立つ.

$$n'' + \sum_{k=1}^{h} (r_k + 1) = n.$$

解析的集合のパラメータ表示を証明する.

4.3.4 代数的集合のパラメータ表示

真の分布が学習モデルによって実現されるパラメータの集合 W_0 を次で定める.

$$W_0 := \{ \ w \ \in \mathbb{R}^{2n} \mid p(x, y|w) = q(x, y)\}.$$

定理 13 (副論文 [26]) (パラメータ表示) 学習モデルを 1 つの入力層, 中間ユニット $H = n$, 1 つの出力ユニットである 3 層ニューラルネットワーク, 真の分布を中間ユニット $H = m$ である 3 層ニューラルネットワーク, 活性化関数を tanh, 真の分布が学習モデルによって実現される解析的集合を W_0 とする.

このとき，$w \in W_0 \subset \mathbb{R}^{2n}$ に対して解析的集合 W_0 のパラメータ表示とパラメータは次の (1), (2), (3) により，$1 \leq i \leq m$, $1 \leq j \leq h$ に対して

$$w = \{\mathbf{a'_1}, \mathbf{b'_1}, \cdots, \mathbf{a'_i}, \mathbf{b'_i}, \cdots, \mathbf{a'_m}, \mathbf{b'_m}, \mathbf{a_0}, \mathbf{b_0}, \mathbf{a_1}, \mathbf{b_1}, \cdots, \mathbf{a_j}, \mathbf{b_j}, \cdots, \mathbf{a_h}, \mathbf{b_h}\}$$

と表される.

(1) $\{\mathbf{a'_i}, \mathbf{b'_i}\}$ のパラメータ表示は $1 \leq i \leq m$ に対して

$$\{\mathbf{a'_i}, \mathbf{b'_i}\} := \{\lambda_i^1 a'_i, b'_i, \lambda_i^2 a'_i, b'_i, \ldots, \lambda_i^{l_i-1} a'_i, b'_i, \left(1 - \lambda_i^1 - \lambda_i^2 \cdots - \lambda_i^{l_i-1}\right) a'_i, b'_i\}.$$

ここで $\{\mathbf{a'_i}, \mathbf{b'_i}\}$ のパラメータは

$$\lambda_i^{l_i-1}.$$

(2) $\{\mathbf{a_0}, \mathbf{b_0}\}$ のパラメータ表示は $n' + 1 \leq k \leq n''$ に対して

$$\{\mathbf{a_0}, \mathbf{b_0}\} := \{a_{n'+1}, b_{n'+1}, \cdots, a_k, b_k, \cdots, a_{n''}, b_{n''}\}.$$

ここで $\{\mathbf{a_0}, \mathbf{b_0}\}$ のパラメータは

$$non - zero\ parameters \quad a_k\ or\ b_k.$$

(3) $\{\mathbf{a_j}, \mathbf{b_j}\}$ のパラメータ表示は $1 \leq j \leq h$ に対して

$$\{\mathbf{a_j}, \mathbf{b_j}\} := \{a_{n''_{j-1}+1}, b_{n''+j}, a_{n''_{j-1}+2}, b_{n''+j}, \ldots, a_{n''_{j-1}+r_j}, b_{n''+j},$$
$$-a_{n''_{j-1}+1} - a_{n''_{j-1}+2} \cdots - a_{n''_{j-1}+r_i}, b_{n''+j}\}.$$

ここで $\{\mathbf{a_j}, \mathbf{b_j}\}$ のパラメータは

$$a_{n''_{j-1}+1}, a_{n''_{j-1}+2}, \ldots, a_{n''_{j-1}+r_j}, b_{n''+j}.$$

証明 (1) 真の分布が学習モデルによって実現される次の解析的集合を考える.

$$\begin{aligned} W_0 &:= \{w \in \mathbb{R}^{2n} | p(x, y|w) = q(x, y)\} \\ &= \{w \in \mathbb{R}^{2n} | a_1 \tanh(b_1 x) + \cdots + a_n \tanh(b_n x) = a'_1 \tanh(b'_1 x) \\ &\qquad + \cdots + a'_m \tanh(b'_m x)\}. \end{aligned}$$

テイラー展開を用いて次を得る.

$$W_0 = \{\ w \in \mathbb{R}^{2n} \mid \sum_{i=1}^{n} a_i b_i^{2k-1} - \sum_{i=1}^{m} {a'}_i {b'}_i^{2k-1} = 0\,(k \geq 1)\}.$$

$k \geq 1$ に対して，定義方程式が条件 $b_i = b'_k\,(1 \leq k \leq m)$ の下で次のように表される．

$$\sum_{i=1}^{m}\left(a_{n'_{i-1}+1}+a_{n'_{i-1}+2}+\cdots+a_{n'_{i-1}+l_i}\right)b_i^{2k-1}+\sum_{i=n'+1}^{n}a_i b_i^{2k-1}-\sum_{i=1}^{m}a'_i b'^{2k-1}_i=0.$$

$k \geq 1$ に対して次が成り立つ．

$$\sum_{i=1}^{m}\left(a_{n'_{i-1}+1}+a_{n'_{i-1}+2}+\cdots+a_{n'_{i-1}+l_i}-a'_i\right)b'^{2k-1}_i+\sum_{i=n'+1}^{n}a_i b_i^{2k-1}=0.$$

$d = n + m$ と定めると，$1 \leq k \leq d$ 対して補題 3 より次が成り立つ．

$$V(I_{d-1}) =$$
$$\{w \in \mathbb{R}^{2n} \mid \sum_{i=1}^{m}\left(a_{n'_{i-1}+1}+a_{n'_{i-1}+2}+\cdots+a_{n'_{i-1}+l_i}-a'_i\right)b'^{2k-1}_i+\sum_{i=n'+1}^{n}a_i b_i^{2k-1}=0\}.$$

補題 5 より消去イデアル I_{d-1} について次が成り立つ．

$$\left(a_{n'_{i-1}+1}+a_{n'_{i-1}+2}+\cdots+a_{n'_{i-1}+l_i}-a'_i\right)b'_i$$
$$\left(b'^2_i-b'^2_1\right)\left(b'^2_i-b'^2_2\right)\cdots\left(b'^2_i-b'^2_m\right)\left(b'^2_i-b^2_{n'+1}\right)$$
$$\left(b'^2_i-b^2_{n'+2}\right)\cdots\left(b'^2_i-b^2_n\right)=0.$$

$1 \leq i \leq m$ に対して方程式を減らすことができる．

$$a_{n'_{i-1}+1}+a_{n'_{i-1}+2}+\cdots+a_{n'_{i-1}+l_i}-a'_i=0.$$

したがって $1 \leq i \leq m$，$\lambda_i^l\,(1 \leq l \leq l_i - 1)$ に対して $\{\mathbf{a'_i}\}$ のパラメータ表示は次で表される．

$$\{\mathbf{a'_i}\} := \{\lambda_i^1 a'_i, \lambda_i^2 a'_i, \ldots, \lambda_i^{l_i-1}a'_i, \left(1 - \lambda_i^1 - \lambda_i^2 \cdots - \lambda_i^{l_i-1}\right)a'_i\}.$$

また，$1 \leq i \leq m$ に対して $\{\mathbf{a'_i}, \mathbf{b'_i}\}$ のパラメータは次で表される．

$$\lambda_i^{l_i-1}.$$

(2) 中間ユニット数が $H = n - n'$ である学習モデルと中間ユニット数が $H_0 = 0$ である真の分布を得る．

解析的集合が真の分布が学習モデルによって実現される場合を考える．

$$W_1 := \{\ w\ \in \mathbb{R}^{2(n-n\prime)}\ \mid a_{n'+1}\tanh(b_{n'+1}x)+\cdots\ +a_n\tanh(b_n x)=0\}.$$

$d' = n - n'$ と定める.

補題 3 より次が成り立つ.

$$V(I_{d'-1}) = \{\ w \in \mathbb{R}^{2(n-n')} | \sum_{i=n'+1}^{n} a_i b_i^{2k-1} = 0\ \ (1 \leq k \leq d')\}.$$

補題 5 より消去イデアル I_{d-1} について次が成り立つ.

$$a_{n'+j} b_{n'+j} \left(b_{n'+j}^2 - b_{n'+1}^2\right) \left(b_{n'+j}^2 - b_{n'+2}^2\right) \cdots \left(b_{n'+j}^2 - b_n^2\right) = 0.$$

次の条件が成り立つ.

$$a_i b_i = 0\,(n'+1 \leq i \leq n''), a_i b_i \neq 0\,(n''+1 \leq i \leq n).$$

また，$n'+1 \leq i \leq n''$ に対して $\{\mathbf{a_0}, \mathbf{b_0}\}$ のパラメータは次で表される.

$$non - zero\ parameters \quad a_i\ or\ b_i.$$

(3) 中間ユニット数が $H = n - n''$ である学習モデルと中間ユニット数が $H_0 = 0$ である真の分布を得る.

解析的集合が真の分布が学習モデルによって実現される場合を考える.

$$W_2 := \{\ w\ \ \in \mathbb{R}^{2(n-n'')}\ \ | a_{n''+1} \tanh\left(b_{n''+1} x\right) + \cdots + a_n \tanh\left(b_n x\right) = 0\}.$$

$d'' = n - n''$ と定める.

テイラー展開を活性化関数に用いて，次を得る.

$$W_2 := \{\ w\ \ \in \mathbb{R}^{2(n-n'')}\ \ | \sum_{i=n''+1}^{n} a_i b_i^{2k-1} = 0\,(k \geq 1)\}.$$

$k\ \geq 1$ に対して，定義方程式が $b_i = b_k\,(n''+1 \leq k \leq n''+h)$ の条件の下で次が成り立つ.

$$\sum_{i=1}^{h} \left(a_{n''_{i-1}+1} + a_{n''_{i-1}+2} + \cdots + a_{n''_{i-1}+r_i}\right) b_{n''+i}^{2k-1} = 0.$$

補題 3 より $1\ \leq i\ \leq h$，$1 \leq k \leq d$ に対して次が成り立つ.

$$V\left(I_{d''-1}\right) = \{\ w\ \ \in \mathbb{R}^{2(n-n'')}\ | \sum_{i=1}^{h} \left(a_{n''_{i-1}+1} + a_{n''_{i-1}+2} + \cdots + a_{n''_{i-1}+r_i}\right) b_{n''+i}^{2k-1} = 0\}.$$

補題 5 より消去イデアル $I_{d''-1}$ について次が成り立つ.

$$\left(a_{n''_{i-1}+1}+a_{n''_{i-1}+2}+\cdots+a_{n''_{i-1}+r_i}\right)b_{n''+i}$$
$$\left(b^2_{n''+i}-b^2_{n''+1}\right)\cdots\left(b^2_{n''+i}-b^2_{n''+h}\right)=0.$$

方程式を減らすことができる.

$$a_{n''_{i-1}+1}+a_{n''_{i-1}+2}+\cdots+a_{n''_{i-1}+r_i}=0.$$

したがって $1\ \leq j\ \leq h$, に対して $\{\mathbf{a_j}\}$ のパラメータ表示は次で表される.

$$\{\mathbf{a_j}\}:=\{a_{n''_{j-1}+1},a_{n''_{j-1}+2},\ldots,a_{n''_{j-1}+r_j},-a_{n''_{j-1}+1}a_{n''_{j-1}+2}\cdots-a_{n''_{j-1}+r_i}\}.$$

また，$1\ \leq j\ \leq h$ に対して $\{\mathbf{a_j},\mathbf{b_j}\}$ のパラメータは次で表される.

$$a_{n''_{j-1}+1},a_{n''_{j-1}+2},\ldots,a_{n''_{j-1}+r_j},b_{n''+j}.$$

□

4.3.5 真の分布を実現するパラメータ集合の次元

真の分布が学習モデルによって実現されるパラメータの集合の次元を考える

定理 14 学習モデルを 1 つの入力層，中間ユニット $H=n$，1 つの出力ユニットである 3 層ニューラルネットワーク，真の分布を中間ユニット $H=m$ である 3 層ニューラルネットワーク，活性化関数を tanh，真の分布が学習モデルによって実現される解析的集合を W_0 とする.
このとき，$w\ \in W_0\subset\mathbb{R}^{2n}$ に対して解析的集合 W_0 の次元は次で表される.

$$\dim\ W_0\ =n-m.$$

証明 $\mathbb{R}^{2n}$ を次のように表示する.

$$\mathbb{R}^{2n}=\mathbb{R}^{2n'}\times\mathbb{R}^{2\left(n''-n'\right)}\times\mathbb{R}^{2\left(n-n''\right)}.$$

$1\ \leq i\ \leq m,\ 1\ \leq j\ \leq h$, に対して w_1,w_2,w_3 と置くと次が成り立つ.

$$w_1=\{\mathbf{a'_1},\mathbf{b'_1},\cdots,\mathbf{a'_i},\mathbf{b'_i},\cdots,\mathbf{a'_m},\mathbf{b'_m}\}\ \in\mathbb{R}^{2n'},$$

$$w_2 = \{\mathbf{a_0}, \mathbf{b_0}\} \in \mathbb{R}^{2\left(n\prime\prime - n'\right)},$$

$$w_3 = \{\mathbf{a_1}, \mathbf{b_1}, \cdots, \mathbf{a_j}, \mathbf{b_j}, \cdots, \mathbf{a_h}, \mathbf{b_h}\} \in \mathbb{R}^{2(n-n\prime\prime)}.$$

V_1, V_2, V_3 を定めると次が成り立つ.

$$V_1 = \{\ w_1\ \in \mathbb{R}^{2n'}\ | a_{n'_{i-1}+1} + a_{n'_{i-1}+2} + \cdots + a_{n'_{i-1}+l_i} - a'_i = 0, (1 \leq i \leq m)\},$$

$$V_2 = \{\ w_2\ \in \mathbb{R}^{2\left(n\prime\prime - n'\right)}\ | a_i b_i = 0, (n'+1 \leq i \leq n'')\},$$

$$V_3 = \{\ \ w_3\ \in \mathbb{R}^{2(n-n\prime\prime)}\ | a_{n''_{i-1}+1} + a_{n''_{i-1}+2} + \cdots\ + a_{n''_{i-1}+r_i} = 0, (1 \leq i \leq h)\}.$$

W_0 を次のように表す.

$$W_0 = V_1 \times V_2 \times V_3.$$

解析的集合 W_0 の次元について次のように表す.

$$\dim\ W_0 = \dim\ V_1 + \dim V_2 + dim V_3.$$

次の定義を用いて

$$\sum_{i=1}^{m} l_i = n', z = n'' - n', \sum_{k=1}^{h} (r_k + 1) = n - n''.$$

次が成り立つ.

$$\begin{aligned}\dim\ W_0 &= \sum_{i=0}^{m} (l_i - 1) + 1 \times z + \sum_{k=0}^{h} (r_k + 1) = \sum_{i=0}^{m} l_i - m + z + (n - n'') \\ &= n' - m + (n'' - n') + n - n'' = n - m.\end{aligned}$$

□

4.4 定理の適用例

4.4.1 中間ユニット数 $H = 2 \to H_0 = 0$ の場合

パラメータ表示，次元

例 1 学習モデルを *1* つの入力層，中間ユニット $H = 2$，*1* つの出力ユニットである 3 層ニューラルネットワーク（1 層から第 2 層への重みを b_1 と b_2，第 2 層から第 3 層への重

みを a_1 と a_2，第バイアスをなしとする.)，活性化関数を tanh，真の分布を中間ユニット $H_0 = 0$ である 3 層ニューラルネットワークとする．このとき，学習モデルが真のモデルを実現する場合を考える.

真の分布が学習モデルによって実現されるパラメータの集合を考える．次で表される元に対して

$$w = \{a_1, b_1, a_2, b_2\} \in W_0.$$

W_0 のパラメータは

$$(a_1, 0, 0, b_2)\,,\ (a_1, 0, a_2, 0)\,,\ (0, b_1, 0, b_2)\,,\ (0, b_1, a_2, 0)\,,$$

$$(a_2, \pm b_2, \mp a_2, b_2) \quad \text{(複合同順)}\,.$$

また W_0 の次元は次で表される.

$$\dim\ W_0 = 2.$$

証明　W_0 を真の分布が学習モデルを実現する解析的集合とする.

$$W_0 := \{\ w\ \in \mathbb{R}^4 \mid p(x, y|w) = q(x, y)\} \ = \{\ w\ \in \mathbb{R}^4\ | a_1 \tanh(b_1 x) + a_2 \tanh(b_2 x) = 0\}.$$

Mathematica を用いてグレブナ基底の消去イデアルの基底を計算するために次のように入力する*:*

$$\mathtt{f1} = \mathtt{a_1 b_1} + \mathtt{a_2 b_2}; \mathtt{f2} = \mathtt{a_1 b_1^3} + \mathtt{a_2 b_2^3}.$$

グレブナ基底の a_2 の消去イデアルの基底を計算するために次のように入力する*:*

$$\mathtt{GroebnerBasis}[\{\mathtt{f1}, \mathtt{f2}\}, \{\mathtt{a_1}, \mathtt{b_1}, \mathtt{a_2}, \mathtt{b_2}\}, \{\mathtt{a_2}\}$$

$$\mathtt{MonomialOrder} \to \mathtt{exicographic}]$$

出力は

$$\mathtt{a_1 b_1^3} - \mathtt{a_1 b_1 b_2^2}.$$

因数分解できる.

$$a_1 b_1^3 - a_1 b_1 b_2^2 = a_1 b_1 \left(b_1^2 - b_2^2\right)\ = a_1 b_1 (b_1 + b_2)(b_1 - b_2) = 0.$$

したがって

$$a_1 = 0 \ or \ b_1 = 0 \text{ or } b_2 = \pm b_1.$$

(i) $a_1 = 0$ に対して $a_2 b_2 = 0$. したがって

$$(0, b_1, 0, b_2), (0, \ 0, a_2, b_2).$$

(ii) $b_1 = 0$ に対して $a_2 b_2 = 0$. したがって

$$(a_1, 0, \ 0, b_2), (a_1, 0, a_2, 0).$$

(iii) $b_1 = \pm b_2$ に対して $(\pm a_1 + a_2) b_2 = 0$. したがって

$$\pm a_1 + a_2 = 0 \text{ or } b_2 = 0.$$

$\pm a_1 + a_2 = 0$ に対して

$$(a_1, \pm b_2, \mp a_1, b_2) \quad \text{(複合同順)}.$$

$b_2 = 0$ に対して

$$(a_1, 0, a_2, 0).$$

真の分布が学習モデルによって実現される解析的集合 W_0 の要素 w のパラメータ表示は次で表される.

$$(a_1, 0, \ 0, \ b_2), (a_1, 0, a_2, 0), (0, b_1, 0, b_2), (0, b_1, a_2, 0),$$
$$(a_2, \pm b_2, \mp a_2, b_2) \quad \text{(複合同順)}.$$

ついに $w = \{a_1, b_1, a_2, b_2\} \in W_0$ に対して V_2, V_3 を定めると以下が成り立つ.

$$V_2 = \{ \ w \ \in \mathbb{R}^4 \ | a_i b_i = 0 \, (1 \leq i \leq 2)\},$$
$$V_3 = \{ \ \ w \ \in \mathbb{R}^4 \ | a_1 + a_2 = 0\}.$$

解析的集合 W_0 は次のように表される.

$$W_0 = V_2 \ , \ V_3.$$

解析的集合 W_0 の次元は次のように得られる.

$$\dim W_0 = n - m = 2 - 0 = 2.$$

□

臨界点，特異点

例 2　真の分布が学習モデルによって実現されるパラメータの集合を考える．

次で表される元に対して

$$w = \{a_1, b_1, a_2, b_2\} \in W_0.$$

臨界点の集合は次で表される．

$(a_1, 0, 0, b_2)$, $(a_1, 0, a_2, 0)$, $(0, b_1, 0, b_2)$, $(0, b_1, a_2, 0)$, $(a_1, b_1, -a_1, b_1)$, $(a_1, b_1, a_1, -b_1)$.

W_0 の特異点集合は次で表される．

$$(0, 0, 0, 0),\ (a_1, 0, 0, 0),\ (0, b_1, 0, 0),\ (0, 0, a_2, 0),\ (0, 0, 0, b_2),\ (a_1, 0, a_2, 0).$$

証明　次に，*Mathematica* を用いて，代数的集合について計算を行う．

$f = (a_1 b_1 + a_2 b_2)^2 + (a_1 {b_1}^3 + a_2 {b_2}^3)^2$ に対して，次の方程式を満たす解を求める．

$$\frac{\partial f}{\partial a} = 2b_1(a_1 b_1(1 + {b_1}^4) + a_2 b_2(1 + {b_1}^2 {b_2}^2)) = 0,$$

$$\frac{\partial f}{\partial b_1} = 2a_1(a_1 b_1(1 + 3b^4) + a_2 d(1 + 3{b_1}^2 {b_2}^2)) = 0,$$

$$\frac{\partial f}{\partial a_2} = 2d(a_1 b_1(1 + {b_1}^2 d^2) + a_2 b_2(1 + {b_2}^4)) = 0,$$

$$\frac{\partial f}{\partial d} = 2a_2(a_1 b_1(1 + 3{b_1}^2 d^2) + a_2 b_2(1 + 3{b_2}^4)) = 0.$$

臨界点の集合は次で表される．

$(a_1, 0, 0, d)$, $(a_1, 0, a_2, 0)$, $(0, b_1, 0, b_2)$, $(0, b_1, a_2, 0)$, $(a_1 b_1, -a_1, b_1)$, $(a_1, b_1, a_1, -b_1)$.

次にヤコビ行列の階数を考える．$A = \begin{pmatrix} b_1 & a_1 & b_2 & a_2 \\ {b_1}^3 & 3a_1{b_1}^2 & {b_2}^3 & 3a_2{b_2}^2 \end{pmatrix}$ とし，

$(a, 0, 0, d)$ のとき，$\mathrm{rank} A = \begin{cases} 2 & (a_1 \neq 0, b_2 \neq 0) \\ 1 & (a_1 = 0, b_2 \neq 0 \text{ 又は } b_2 = 0, a_1 \neq 0), \\ 0 & (a_1 = 0, b_2 = 0) \end{cases}$

$$(a_1,0,c,0)\text{ のとき, }\mathrm{rank}A=\begin{cases}1 & (a_1\neq 0,a_2\neq 0)\\ 1 & (a_1=0,a_2\neq 0\text{ 又は }a_2=0,a_1\neq 0),\\ 0 & (a_1=0,a_2=0)\end{cases}$$

$$(0,b,0,d)\text{ のとき, }\mathrm{rank}A=\begin{cases}2 & (b_1\neq 0,b_2\neq 0)\\ 1 & (b_1=0,b_2\neq 0\text{ 又は }d=0,b_1\neq 0),\\ 0 & (b_1=0,b_2=0)\end{cases}$$

$$(0,b,c,0)\text{ のとき, }\mathrm{rank}A=\begin{cases}2 & (b_1\neq 0,a_2\neq 0)\\ 1 & (b_1=0,a_2\neq 0\text{ 又は }a_2=0,b_1\neq 0),\\ 0 & (b_1=0,a_2=0)\end{cases}$$

であり, $(a_1,b_1,-a_1,b_1),(a_1,b_1,a_1,-b_1)$ のとき, $\mathrm{rank}A=2$ である.
よって特異点集合は次で表される.
$(0,0,0,0)$, $(a_1,0,0,0)$, $(0,b_1,0,0)$, $(0,0,a_2,0)$, $(0,0,0,b_2)$, $(a_1,0,a_2,0)$. □

4.4.2 中間ユニット数 $H=2\to H_0=1$ の場合

例 3 学習モデルを 1 つの入力層, 中間ユニット $H=2$, 1 つの出力ユニットである 3 層ニューラルネットワーク（第 1 層から第 2 層への重みを b_1 と b_2, 第 2 層から第 3 層への重みを a_1 と a_2, バイアスをなしとする.）, 活性化関数を tanh, 真の分布を中間ユニット $H_0=1$ である 3 層ニューラルネットワーク（第 1 層から第 2 層への重みを ${b'}_1$, 第 2 層から第 3 層への重みを ${a'}_1$, バイアスをなしとする.）とする.

このとき, 学習モデルが真のモデルを実現する場合を考える.

真の分布が学習モデルによって実現されるパラメータの集合を考える.
次で表される元に対して

$$w=\{a_1,b_1,a_2,b_2\}\in W_0.$$

W_0 のパラメータは

$$\left(\pm a'_1,\pm b'_1,a_2,0\right),\left(\pm a'_1,\pm b'_1,0,b_2\right)\text{ (複合同順)},\left(\lambda a'_1,\pm b'_1,\pm\left(1\mp\lambda\right)a'_1,\pm b'_1\right).$$

(λ の第 2, 第 3 成分と第 4, 第 3 成分は複合同順で第 2 成分と第 4 成分は複合任意)

また W_0 の次元は次で表される.

$$\dim W_0 = 2 - 1 = 1.$$

証明 W_0 を真の分布が学習モデルを実現する解析的集合とする.

$$\begin{aligned} W_0 &:= \{\ w\ \in \mathbb{R}^4 \mid p(x,y|w) = q(x,y)\} \\ &= \{\ w\ \in \mathbb{R}^4 \mid a_1 \tanh(b_1 x) + a_2 \tanh(b_2 x) = a_1' \tanh(b_1' x)\}. \end{aligned}$$

活性化関数にテイラー展開を用いて，次を得る.

$$W_0 = \{\, w \in \mathbb{R}^4 \mid a_1 b_1 + a_2 b_2 - a\prime_1 b_1' = a_1 b_1^3 + a_2 b_2^3 - a_1' {b'}_1^3 = a_1 b_1^5 + a_2 b_2^5 - a_1' {b'}_1^5 = 0\}.$$

$n = 2$ に対して

$$a_1 b_1 + a_2 b_2 - a_1' b_1' = a_1 b_1^3 + a_2 b_2^3 - a_1' {b'}_1^3 = a_1 b_1^5 + a_2 b_2^5 - a_1' {b'}_1^5 = 0.$$

Mathematica を用いてグレブナ基底の消去イデアルの基底を計算するために次のように入力する*:*

$$\mathtt{f1} = \mathtt{a_1 b_1} + \mathtt{a_2 b_2} - \mathtt{a_1' b_1'};\ \mathtt{f2} = \mathtt{a_1 b_1^3} + \mathtt{a_2 b_2^3} - \mathtt{a_1' b'^3_1};\ \mathtt{f3} = \mathtt{a_1 b_1^5} + \mathtt{a_2 b_2^5} - \mathtt{a_1' b'^5_1};$$

グレブナ基底の a_1, a_2 の消去イデアルの基底を計算するために次のように入力する*:*

$$\mathtt{GroebnerBasis}[\{\mathtt{f1},\ \mathtt{f2},\ \mathtt{f3}\}, \{\mathtt{a_1}, \mathtt{b_1}, \mathtt{a_2}, \mathtt{b_2}, \mathtt{a_1'}, \mathtt{b_1'}\},\ \{\mathtt{a_1}, \mathtt{a_2}\}$$

$$\mathtt{MonomialOrder}\ \rightarrow\ \mathtt{exicographic}];$$

出力は

$$\mathtt{b_1}^2 \mathtt{b_2}^2 \mathtt{a_1'} \mathtt{b_1'} - \mathtt{b_1}^2 \mathtt{a_1'} \mathtt{b'^3_1} - \mathtt{b_2}^2 \mathtt{a_1'} \mathtt{b_1'}^3 + \mathtt{a_1'} \mathtt{b'^5_1}.$$

次のように因数分解できる.

$$a_1' b_1' \left({b'}_1^3 - {b_1}^2 \right) \left({b'}_1^3 - {b_2}^2 \right) = 0.$$

したがって

$$b_1 = \pm b_1'\ or\ b_2 = \pm b_1'.$$

方程式の定義より (a_i, b_i) を並び替えて次を得る.

$$b_1 = \pm b_1'.$$

(i) $b_1 = b_1'$ に対して *Mathematica* を用いてグレブナ基底の a_1 の消去イデアルの基底を計算するため次のように入力する*:*

$$\mathtt{f1} = (\mathtt{a_1} - \mathtt{a_1'})\,\mathtt{b_1'} + \mathtt{a_1'b_2};\ \mathtt{f2} = (\mathtt{a_1} - \mathtt{a_1'})\,\mathtt{b'}_1^3 + \mathtt{a_1'b_2}^3;\ \mathtt{f3} = (\mathtt{a_1} - \mathtt{a_1'})\,\mathtt{b'}_1^5 + \mathtt{a_1'b_2}^5;$$

グレブナ基底の a_1 の消去イデアルの基底を計算するため次のように入力する*:*

$$\mathtt{GroebnerBasis}[\{\mathtt{f1},\ \mathtt{f2},\ \mathtt{f3}\}, \{\mathtt{a_1}, \mathtt{a_2}, \mathtt{b_2}, \mathtt{a_1'}, \mathtt{b_1'}\},\ \{\mathtt{a_1}\}$$

$$\mathtt{MonomialOrder}\ \rightarrow\ \mathtt{exicographic}];$$

出力は

$$\mathtt{a_2b_2}^3 - \mathtt{a_2b_2b'}_1^3.$$

次のように因数分解できる.

$$a_2b_2\left(b_1'^3 - {b_2}^2\right) = 0.$$

(ii) $b_1 = -b_1'$ に対して *Mathematica* を用いてグレブナ基底の消去イデアルの基底を計算するため次のように入力する*:*

$$\mathtt{f1} = (-\mathtt{a_1} - \mathtt{a_1'})\,\mathtt{b_1'} + \mathtt{a_1'b_2};\ \mathtt{f2} = (-\mathtt{a_1} - \mathtt{a_1'})\,\mathtt{b'}_1^3 + \mathtt{a_1'b_2}^3;\ \mathtt{f3} = (-\mathtt{a_1} - \mathtt{a_1'})\,\mathtt{b'}_1^5 + \mathtt{a_1'b_2}^5$$

グレブナ基底の a_1 の消去イデアルの基底を計算するため次のように入力する*:*

$$\mathtt{GroebnerBasis}[\{\mathtt{f1},\ \mathtt{f2},\ \mathtt{f3}\}, \{\mathtt{a_1}, \mathtt{a_2}, \mathtt{b_2}, \mathtt{a_1'}, \mathtt{b_1'}\},\ \{\mathtt{a_1}\};$$

$$\mathtt{MonomialOrder}\ \rightarrow\ \mathtt{exicographic}].$$

出力は

$$\mathtt{a_2b_2}^3 - \mathtt{a_2b_2b_1'}^2.$$

次のように因数分解できる.

$$a_2b_2\left(b_1'^2 - {b_2}^2\right) = 0.$$

(i) と (ii) から $b_1 = \pm b_1'$ に対して次が成り立つ.

$$a_2b_2\left(b_1'^2 - {b_2}^2\right) = 0.$$

したがって

$$a_2 = 0 \text{ or } b_2 = 0 \text{ or } b_2 = \pm b_1'.$$

$a_2 = 0\ or\ b_2 = 0$ に対して

$$(\pm a_1 - a_1')\,b_1' = 0.$$

よって

$$\left(\pm a'_1, \pm b'_1, 0, b_2\right), \left(\pm a'_1, \pm b'_1, a_2, 0\right) \text{(複合同順)}.$$

$b_2 = \pm b'_1$ に対して

$$\left(\pm a'_1 \pm b'_1 - a_2\right) a'_1 = 0 \text{(複合同順)}.$$

$\pm a'_1 \pm b'_1 = a_2$ に対して

$$\left(\lambda a'_1, \pm b'_1, \pm\left(1 \mp \lambda\right) a'_1, \pm b'_1\right)$$

(λ の第 2, 第 3 成分と第 4, 第 3 成分は複合同順で第 2 成分と第 4 成分は複合任意.)

真の分布が学習モデルによって実現される解析的集合 W_0 の要素 w のパラメータ表示は次で表される.

$$\left(\pm a'_1, \pm b'_1, a_2, 0\right), \left(\pm a'_1, \pm b'_1, 0, b_2\right) \text{(複合同順)},$$

$$\left(\lambda a'_1, \pm b'_1, \pm\left(1 \mp \lambda\right) a'_1, \pm b'_1\right)$$

(λ の第 2, 第 3 成分と第 4, 第 3 成分は複合同順で第 2 成分と第 4 成分は複合任意.)

$w = \{a_1, b_1, a_2, b_2\} \in W_0$ に対して V_1, V'_1, V_2 を定めると以下が成り立つ.

$$V_1 = \{\ w\ \in \mathbb{R}^4\ | a_1 + a_2 - a'_1 = 0\},$$

$$V'_1 = \{\ w\ \in \mathbb{R}^2\ | a_1 - a'_1 = 0\},$$

$$V_2 = \{\ w\ \in \mathbb{R}^2\ | a_1 b_1 = 0\}.$$

解析的集合 W_0 は次のように表される.

$$W_0 = V_1\ ,\ \ V'_1\ \times\ V_2.$$

解析的集合 W_0 の次元は次のように得られる.

$$\dim\ W_0 = n - m = 2 - 1 = 1.$$

□

第 5 章

ニューラルネットワークの作成

5.1　特異領域

5.1.1　同値類による商空間と特異構造

甘利 ([9]) による特異構造について以下で説明する.

パラメータ $\theta = (w_1, w_2, \cdots, w_d)$，出力関数 $y = f(x, \theta)$ とする．パラメータ θ の空間を W，出力関数の空間を S とすると，$f(x, \theta)$ は W から S への写像である．2 つの異なるパラメータ θ と θ' が $f(x, \theta) = f(x, \theta')$ となるとき，2 つのパラメータは同値であるといい，$\theta \approx \theta'$ と書く．$\approx$ は同値関係を定め，出力関数の空間 S は W を同値類で割った商空間 $W/\approx$ である．同値な点は W の中に広く散らばっており，S 上では 1 点に縮むから S では特異点となる ([9]).

中間ユニット数 $H = 1 \to H_0 = 0$ の場合

学習モデル $a \tanh(bx)$ について，$a = 0$ のとき, b が free を満たす点をすべて 1 点に縮めた特異構造 1 を図 5.1 に示す ([18]).

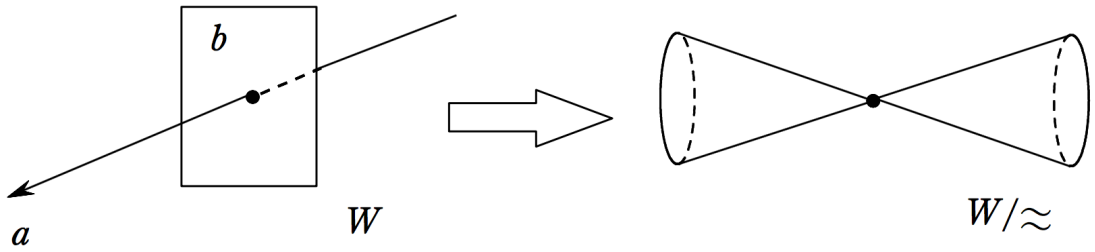


図 5.1　特異構造 1

中間ユニット数 $H = 2 \to H_0 = 1$ の場合

学習モデル $p(x|w) = a\tanh(bx) + c\tanh(dx)$ について, $a = 0$ のとき, b が free を満たす点をすべて 1 点に縮めた特異構造 2 を図 5.2 に表す ([18]).

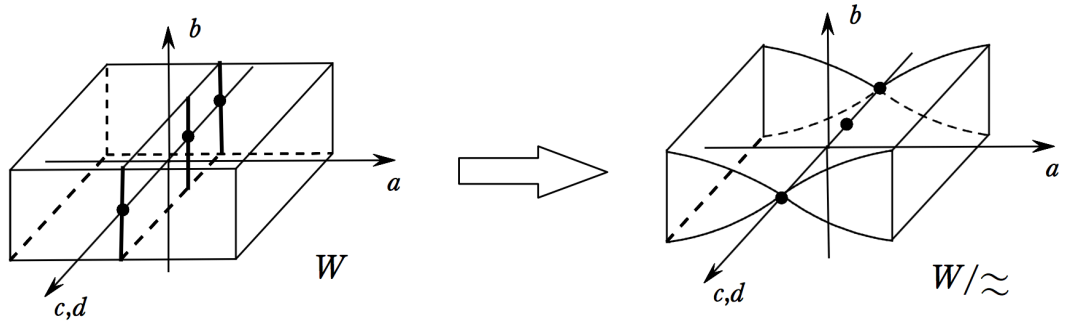


図 5.2　特異構造 2

5.1.2　特異領域

2 つの特異領域 Overlap singularity と Elimination singularity の定義を行う.

定義 44 ([3]) (訓練データ, テストデータ) $\mathbb{R}^1$ 上の一様分布に従う確率変数 X を入力とする. 平均 0, 標準偏差 σ の $\mathbb{R}^1$ 上の正規分布に従う確率変数 Z を雑音とする. $\theta_{\mathbf{0}} = (w_{11}^*, w_{12}^*, w_{21}^*, w_{22}^*, w_3^*, w_4^*) \in \mathbb{R}^6$ に対して, 次で定まる $\mathbb{R}^1$ 上の確率変数 Y を訓練データ, テストデータとする.

$$Y := f(x, \theta_0) + Z = w_3^* \tanh(w_{11}^* x + w_{21}^*) + w_4^* \tanh(w_{12}^* x + w_{22}^*) + Z.$$

パラメータ $\theta = (\mathbf{w_1}, \mathbf{w_2}, w_3, w_4) \in \mathbb{R}^6$, $\mathbb{R}^1$ への関数 $f(x, \theta)$ に対して, 次で定まる $\mathbb{R}^1$ 上の確率変数 Y を関数近似モデルとする ([3]).

$$\begin{aligned} Y &:= f(x, \theta) + Z = w_3 \phi(\mathbf{x}, \mathbf{w_1}) + w_4 \phi(\mathbf{x}, \mathbf{w_2}) + Z \\ &= w_3 \tanh(\mathbf{w_1}^{\mathrm{T}} \mathbf{x}) + w_4 \tanh(\mathbf{w_2}^{\mathrm{T}} \mathbf{x}) + Z \\ &= w_3 \tanh(w_{11} x + w_{21}) + w_4 \tanh(w_{12} x + w_{22}) + Z. \end{aligned}$$

ここで, $\mathbf{w_1} = (w_{11}, w_{12})$, $\mathbf{w_2} = (w_{21}, w_{22})$, $\mathbf{x} = (x, 1)$ とする.

次に関数近似モデル Y が従う条件付き確率を次で定め，学習モデルとする．

$$p(y|x,\theta) := \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{|y-f(x,\theta)|^2}{2\sigma^2}\right).$$

関数近似モデル Y が従う条件付き確率を次で定め，真の分布とする．

$$q(y|x) := \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{|y-f(x,\theta_0)|^2}{2\sigma^2}\right).$$

定義 45 ([9]) (Overlap singularity, Elimination singularity) $\theta = (\mathbf{w_1}, \mathbf{w_2}, w_3, w_4) \in \mathbb{R}^6$ に対して，次で定める $\mathbb{R}^6$ の部分集合を Overlap singularity という．

$$R_0 := \{\theta \in \mathbb{R}^6 | \mathbf{w_1} = \mathbf{w_2}\}.$$

次で定める $\mathbb{R}^6$ の部分集合を Elimination singularity という．

$$R_1 := \{\theta \in \mathbb{R}^6 | w_3 = 0\} \cup \{\theta \in \mathbb{R}^6 | w_4 = 0\}.$$

$f(x,\theta)$ は Overlap singularity では $f(x,\theta) = (w_3+w_4)\phi(\mathbf{x}, \mathbf{w_1})$ と表され, Elimination singularity では $f(x,\theta) = w_4\phi(\mathbf{x}, \mathbf{w_2})$, $w_3\phi(\mathbf{x}, \mathbf{w_1})$ と表される．

5.2 座標変換

座標系 $\theta = (\mathbf{w_1}, \mathbf{w_2}, w_3, w_4)$ から座標系 $\xi = (\mathbf{a}, b, \mathbf{v}, w)$ への座標変換を行う．パラメータ $(\mathbf{a}, b)$ の挙動に関する甘利による研究成果 ([9], [12], [13], [14]) を以下で説明する．

5.2.1 パラメータの座標変換

座標系 $\theta = (\mathbf{w_1}, \mathbf{w_2}, w_3, w_4)$ から座標系 $\xi = (\mathbf{a}, b, \mathbf{v}, w)$ への座標変換を次で定める ([9]).

$$\mathbf{a} = \mathbf{w_2} - \mathbf{w_1}, \ b = \frac{w_3 - w_4}{w_3 + w_4},$$
$$\mathbf{v} = \frac{w_3\mathbf{w_1} + w_4\mathbf{w_2}}{w_3 + w_4}, \ w = w_3 + w_4.$$

このとき座標系 θ は座標系 ξ を用いて次で表される ([9]).

$$\mathbf{w_1} = \mathbf{v} + \frac{1}{2}\mathbf{a}(b-1),\ \mathbf{w_2} = \mathbf{v} + \frac{1}{2}\mathbf{a}(b+1),$$
$$w_3 = \frac{1}{2}w(1+b),\ w_4 = \frac{1}{2}w(1-b).$$

定義 46 ([9]) (Overlap singularity, Elimination singularity) $\xi = (\mathbf{a},\ b,\ \mathbf{v},\ w)$ に対して，次で定める $\mathbb{R}^6$ の部分集合を Overlap singularity という．

$$R_0 := \{\xi \in \mathbb{R}^6 | \mathbf{a} = \mathbf{0}\}.$$

次で定める $\mathbb{R}^6$ の部分集合を Elimination singularity という．

$$R_1 := \{\xi \in \mathbb{R}^6 | b = \pm 1\}.$$

5.2.2 Fast submanifold と Slow submanifold

$y = f(x, \theta_0) + Z$ に対して，誤差関数 $l(y, \mathbf{x}, \theta) := \frac{1}{2}(y - f(\mathbf{x}, \theta))^2$ を定める．このとき学習係数 η に対してパラメータ θ を次のように更新することを学習という ([12]).

$$\theta(t+1) - \theta(t) := -\eta \frac{\partial l(y_t, \mathbf{x_t}, \theta_t)}{\partial \theta}.$$

定義 47 ([12]) (座標系θの学習方程式) $\theta = (\mathbf{w_1},\ \mathbf{w_2},\ w_3,\ w_4)$，誤差関数 $l(y, \mathbf{x}, \theta)$ に対して，学習方程式を次で定める．

$$\dot{\theta}(t) := -\eta \left\langle \frac{\partial l(y, \mathbf{x}, \theta)}{\partial \theta} \right\rangle.$$

ここで $\left\langle \frac{\partial l(y,\mathbf{x},\theta)}{\partial \theta} \right\rangle$ を次のように定める．

$$\left\langle \frac{\partial l(y, \mathbf{x}, \theta)}{\partial \theta} \right\rangle := \int \frac{\partial l(y, \mathbf{x}, \theta)}{\partial \theta} q(y|\mathbf{x}) dy d\mathbf{x}.$$

定義 48 ([12]) (座標系ξの学習方程式) $\xi = (\mathbf{a},\ b,\ \mathbf{v},\ w)$，誤差関数 $l(y, \mathbf{x}, \xi)$ に対して，学習方程式を次で定める．

$$\dot{\xi} := -\eta \frac{\partial \xi}{\partial \theta^{\mathrm{T}}} \left(\frac{\partial \xi}{\partial \theta^{\mathrm{T}}} \right)^{\mathrm{T}} \left\langle \frac{\partial l(y, \mathbf{x}, \xi)}{\partial \xi} \right\rangle .$$

誤差 $e(y, \mathbf{x}, \xi) := y - f(\mathbf{x}, \xi) + Z$ に対して，誤差関数 $l(\xi)$ の勾配について次が成り立つ ([12]).

$$\begin{aligned} l_{\mathbf{v}}(\xi) &= w \left\langle e(y, \mathbf{x}, \xi) \frac{\partial \phi(\mathbf{x}, \mathbf{v})}{\partial \mathbf{v}} \right\rangle + \frac{1}{8} w (1 - z^2) Q(\mathbf{v}, \mathbf{a}) + O(\mathbf{a}^3), \\ l_w(\xi) &= \langle e(y, \mathbf{x}, \xi) \phi(\mathbf{x}, \mathbf{v}) \rangle + \frac{1}{8} (1 - z^2) \left\langle e(y, \mathbf{x}, \xi) \mathbf{a}^{\mathrm{T}} \frac{\partial^2 \phi(\mathbf{x}, \mathbf{v})}{\partial \mathbf{v} \mathbf{v}^{\mathrm{T}}} \mathbf{a} \right\rangle + O(\mathbf{a}^3). \end{aligned}$$

ここで $Q(v, \mathbf{a})$ を次のように定める．

$$Q(v, \mathbf{a}) := \left\langle e(y, \mathbf{x}, \xi) \frac{\partial}{\partial \mathbf{v}} \left(\mathbf{a}^{\mathrm{T}} \frac{\partial^2 \phi(\mathbf{x}, \mathbf{v})}{\partial \mathbf{v} \mathbf{v}^{\mathrm{T}}} \mathbf{a} \right) \right\rangle .$$

$\mathbf{a} \approx \mathbf{0}$ のとき $l_{\mathbf{v}}$, l_w は $O(1)$ のオーダーであるので $l_{\mathbf{v}}$, l_w の変化は速く，パラメータ $(\mathbf{v},\ w)$ は，$l_{\mathbf{v}}(\xi) = \mathbf{l}_{\mathbf{w}}(\xi) = \mathbf{0}$ で定まる部分多様体に速く変化する．この部分多様体を Fast submanifold という ([13]).

誤差関数 $l(\xi)$ の勾配について次が成り立つ ([12]).

$$l_{\mathbf{a}}(\xi) = \frac{1}{4} w (1 - z^2) \left\langle e(y, \mathbf{x}, \xi) \mathbf{a} \frac{\partial^2 \phi(\mathbf{x}, \mathbf{v})}{\partial \mathbf{v} \mathbf{v}^{\mathrm{T}}} \right\rangle + \frac{1}{24} w z (1 - z^2) \left\langle e(y, \mathbf{x}, \xi) \frac{\partial D(x, v, \mathbf{a})}{\partial \mathbf{a}} \right\rangle + O(\mathbf{a}^3),$$

$$l_b(\xi) = -\frac{1}{4} w z \left\langle e(y, \mathbf{x}, \xi) \mathbf{a}^{\mathrm{T}} \frac{\partial^2 \phi(\mathbf{x}, \mathbf{v})}{\partial \mathbf{v} \mathbf{v}^{\mathrm{T}}} \mathbf{a} \right\rangle + O(\mathbf{a}^3).$$

ここで $D(\mathbf{x}, \mathbf{v}, \mathbf{a})$ を次のように定める．

$$D(\mathbf{x}, \mathbf{v}, \mathbf{a}) := \sum_{i,\ j,\ k} \frac{\partial^3 \phi(\mathbf{x}, \mathbf{v})}{\partial v_i v_j v_k} a_i a_j a_k.$$

$\mathbf{a} \approx \mathbf{0}$ のとき $l_{\mathbf{a}}$ は $O(\mathbf{a})$ のオーダー，l_b は $O(\mathbf{a}^2)$ のオーダーであるので $l_{\mathbf{a}}$, l_b の変化は遅い．パラメータ $(\mathbf{a},\ b)$ はこの部分多様体の中で遅く変化する．この部分多様体を Slow submanifolds という ([13]).

ここでパラメータ $(\mathbf{v},\ w)$ を最適解 $(\mathbf{v}^*, w^*)$ に固定して，パラメータ $(\mathbf{a},\ b)$ の変化を分析する ([14]).

5.3 Mathematica を用いたニューラルネットワークの作成

ニューラルネットワークを Mathematica の NetGraph の機能を用いて構成する ([27], [28]).

Mathematica を用いて $F1,\ F2,\ elem0,\ elem1,\ elem2,\ elem3$ を定義するため次のように入力する:

```
F1[a_] := NetInsertSharedArrays[NetChain
[{LinearLayer[1,"Weights"->a,"Biases"->None]}],"Linear1"];
F2[b_] := NetInsertSharedArrays[NetChain
[{LinearLayer[1,"Weights"->b,"Biases"->None]}],"Linear2"];
elem0 := ElementwiseLayer[#*(1/2)&];
elem1 := ElementwiseLayer[#*(-1)&];
elem2[v_] := ElementwiseLayer[#*(v)&];
elem3[w_] := ElementwiseLayer[#*(w)&];
```

5.3.1 net11, net12 の作成

初めに，条件 $w_1 x = \left(v + \frac{1}{2}(b-1)a\right) x$ を表現するため，次のように入力する:

```
net11[a_, b_, v_] := NetGraph[{elem0, elem1, F1[a], F2[b], elem2[v],
TotalLayer[]}, {NetPort["Input"]->1, 1->3->4, 3->2, {4, 2, 5}->6}]
```

次の条件 $w_3 \tanh(x) = \frac{1}{2} w(b+1) \tanh(x)$ を表現するため，次のように入力する:

```
net12[a_, b_, w_] := NetGraph[{Tanh, elem0, elem3[w], F2[b], TotalLayer[]},
{NetPort["Input"]->1, 1->2->3->4, {3, 4}->5}]
```

$net11$ を図 5.3 の上側に，$net12$ を図 5.3 の下側に示す.

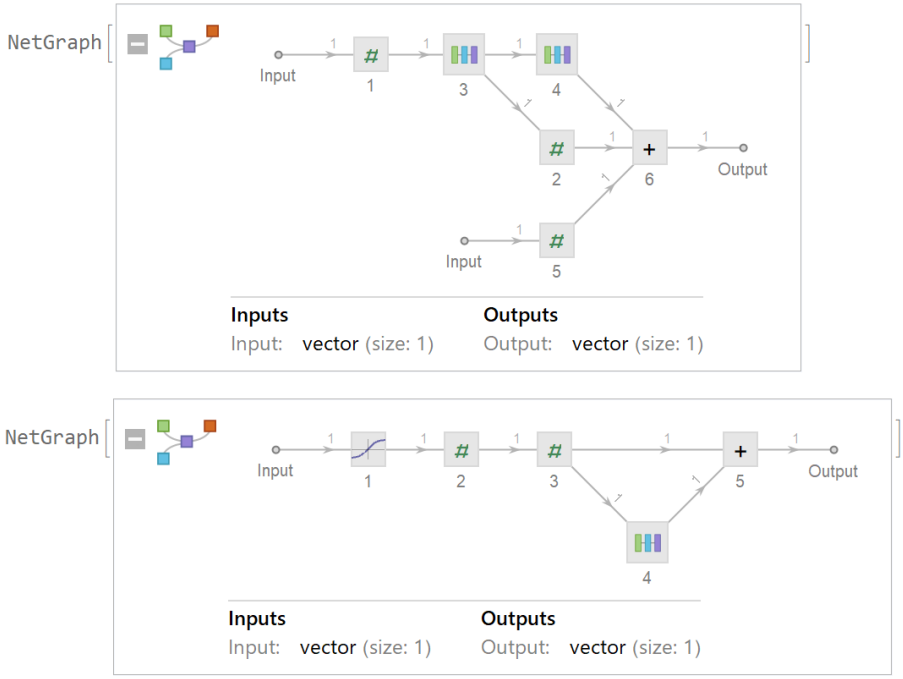


図 5.3　$net11$, $net12$

同様に，条件 $w_2 x = \left(v + \frac{1}{2}(b+1)a\right) x$, $w_4 \tanh(x) = \frac{1}{2} w(-b+1) \tanh(x)$ を表現するため $net21$, $net22$ を定める．

5.3.2　Net1, Net2 の作成

次に，条件 $w_3 \tanh(w_1 x) = \frac{1}{2} w(b+1) \tanh[\left(v + \frac{1}{2}(b-1)a\right) x]$ を表現するため，次のように入力する:

```
net1[a_, b_, v_, w_] := NetGraph[{net11[a, b, v], net12[a, b, w]},
{NetPort["Input"] -> 1, 1 -> 2}]
```

そして $Net1$ を図 5.4 に示す．

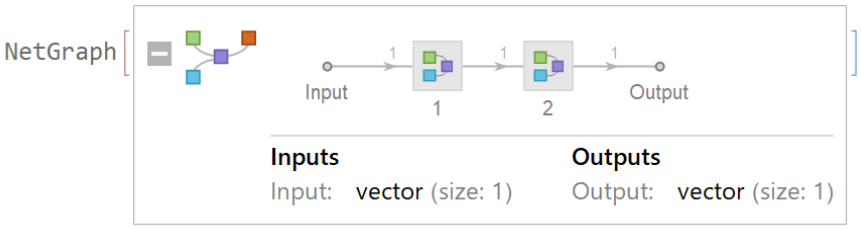


図 5.4　$Net1$

同様に，次の条件 $w_4 \tanh(w_2 x) = \frac{1}{2} w(-b+1) \tanh[\left(v + \frac{1}{2}(b+1)a\right) x]$ を表現する $Net2$ を定める．

5.3.3　parameterNet の作成

最後に，条件 $w_3 \tanh(w_1 x) + w_4 \tanh(w_2 x)$ を表現するため，次のように入力する:

```
parameterNet[a_, b_, v_, w_] := NetGraph[{net1[a, b, v, w], net2[a, b, v, w],
TotalLayer[]}, {NetPort["Input"] -> 1, NetPort["Input"] -> 2, {1, 2} -> 3 ->
NetPort["Output1"]}, "Input" -> enc]
```

そして $parameterNet$ を図 5.5 に示す．

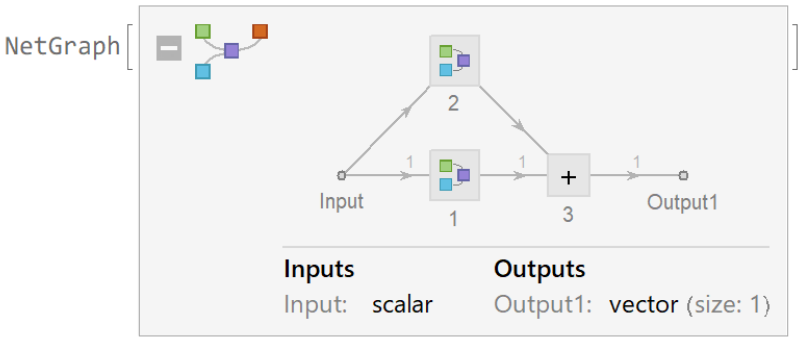


図 5.5　parameterNet

5.3.4 traingNet，traingNet2 の作成

Mathematica ではガウス関数を次で定める．

$$gaussianLikelihood[y, \mu, \sigma] := \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2}).$$

ここで，trainingNet を定義するため，損失関数を対数尤度比関数としてネットグラフを作成する．次のように入力し，図 5.6 に表示される．

```
gaussianLikelihood[y, μ] := PDF[NormalDistribution[μ, 1], y];
trainingNet[a_, b_, v_, w_] := NetGraph[<|"params" -> parameterNet[a, b, v, w], "lhood" ->
ThreadingLayer[gaussianLikelihood], "neglog" -> ElementwiseLayer[-Log[#]&]|>,
{{NetPort["Output"], NetPort["params", "Output1"]} -> "lhood", "lhood" -> "neglog" ->
NetPort["Loss"]}]
```

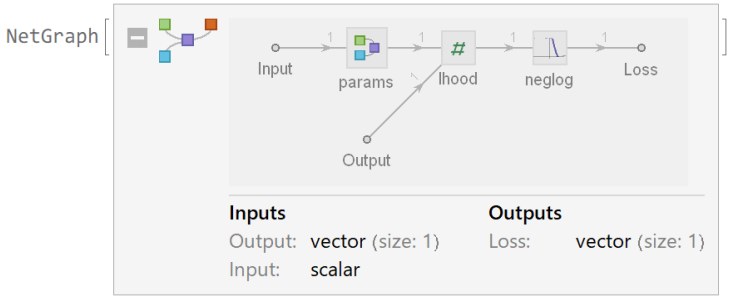


図 5.6 *trainingNet*

損失関数を 2 乗誤差関数として，次のように入力し，図 5.7 に表示される．

```
trainingNet[a_, b_, c_, d_, v_, w_] := NetGraph[<|"params" -> parameterNet[a, b, c, d, v, w],
"loss" -> MeanSquaredLossLayer[]|>, NetPort["Output"], NetPort["params", "Output1"]
-> "loss"]
```

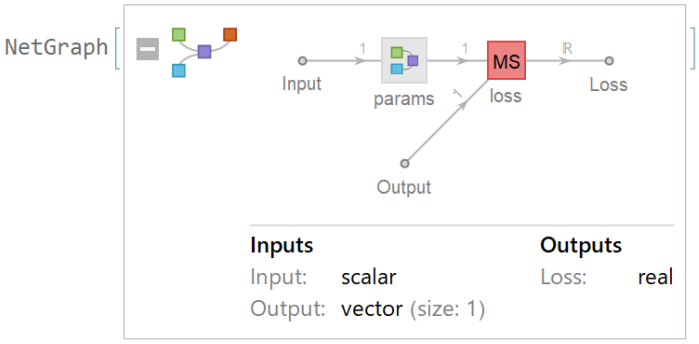


図 5.7　$trainingNet2$(2 乗誤差)

中間ユニット数 $H = 2 \to H_0 = 1$ の場合において，特異構造を定めると，本研究で重要な 2 つの特異領域においてプラトー現象 (学習の停滞) が起こる．その現象を図 5.8 に図示する．

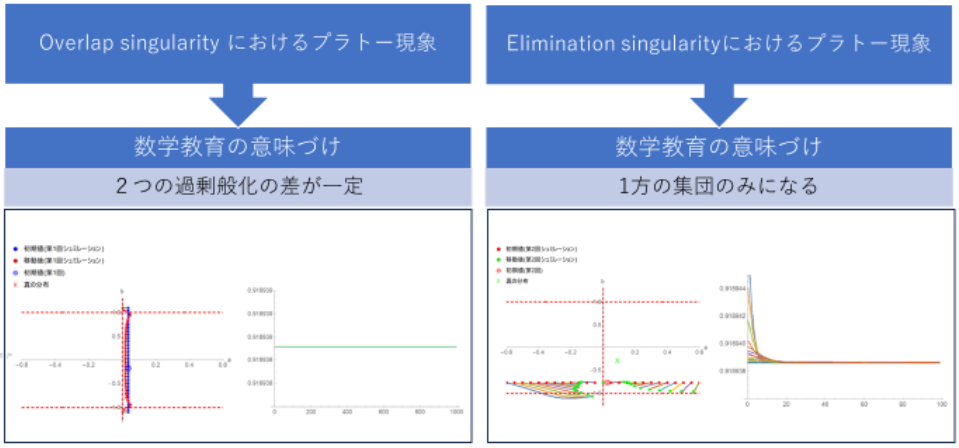


図 5.8　$H = 2 \to H_0 = 1$ におけるプラトー現象

甘利の手法を用いて，第 6 章からニューラルネットワークの重みパラメータを座標変換して収束の遅いパラメータ $\mathbf{a}$, b における学習について考察する．

第 6 章

学習のダイナミクスの分析

中間ユニット数 $H = 2 \to H_0 = 1$ の場合において，甘利 ([9]) による手法を用いて，特異領域の近傍におけるダイナミクスの安定性と学習方程式について述べる．

6.1 特異領域の近くにおける学習のダイナミクスと分類

6.1.1 学習のダイナミクス

R_0 の近傍においてパラメータ $(\mathbf{v}, w)$ の最適解 $(\mathbf{v}^*, w^*)$ に対して，$\xi^* = (\mathbf{v}^*, w^*, \mathbf{0}, b)$ とする．このとき，誤差関数 $l(y, \mathbf{x}, \xi)$ に対して $\xi^* = (\mathbf{v}^*, w^*, \mathbf{0}, b)$ におけるヘッセ行列の値は次のように表される ([9]).

$$\left\langle \frac{\partial^2 l(y, \mathbf{x}, \xi)}{\partial \xi \partial \xi^{\mathrm{T}}} \right\rangle \Bigg|_{\xi=\xi^*} = (1 - b^2) H(\mathbf{v}^*, w^*).$$

ここで $H(\mathbf{v}^*, w^*)$ を次のように定める．

$$H(\mathbf{v}^*, w^*) := \frac{1}{4} w^* \left\langle e(y, \mathbf{x}, \xi) \frac{\partial^2 \phi(\mathbf{x}, \mathbf{v})}{\partial \mathbf{v} \mathbf{v}^{\mathrm{T}}} \right\rangle \Bigg|_{\xi=\xi^*}.$$

特異領域の近くの学習の安定性について次が成り立つ．

定理 15 ([9]) (特異領域の近くの学習の安定性) 誤差関数 $l(y, \mathbf{x}, \xi)$ に対して，真の分布が特異領域にあるとき R_0 は安定である．
真の分布が特異領域上にないとき R_0 の安定性は $H(\mathbf{v}^*, w^*)$ の固有値によって次の 3 つの場合に分かれる．

(1) 固有値が正と負の値を持つ場合: R_0 は不安定である.
(2) 固有値が 2 つとも負の場合: R_0 の中で $|b| < 1$ を満たす部分が安定である.
(3) 固有値が 2 つとも正の場合: R_0 の中で $|b| > 1$ を満たす部分が安定である.

$\tilde{\xi} = (v^*, w^*, a, b)$ に対して，誤差関数 $l(\tilde{\xi})$ の勾配について次が成り立つ ([9]).

$$
\begin{aligned}
l_{\mathbf{v}}(\tilde{\xi}) &= \frac{1}{8} w^* (1-z^2) Q(\mathbf{v}^*, \mathbf{a}) + O(\mathbf{a}^3), \\
l_w(\tilde{\xi}) &= \frac{1}{2} \frac{1-z^2}{w^*} H(\mathbf{v}^*, w^*) \mathbf{a}^2 + O(\mathbf{a}^3), \\
l_{\mathbf{a}}(\tilde{\xi}) &= (1-z^2) H(\mathbf{v}^*, w^*) \mathbf{a} + \\
&\quad + \frac{1}{24} w^* z (1-z^2) \left\langle e(y, \mathbf{x}, \xi) \frac{\partial D(x, \mathbf{v}, \mathbf{a})}{\partial \mathbf{a}} \right\rangle \Big|_{\xi = \tilde{\xi}} + O(\mathbf{a}^3), \\
l_b(\tilde{\xi}) &= -b H(\mathbf{v}^*, w^*) \mathbf{a}^2 + O(\mathbf{a}^3).
\end{aligned}
$$

$l_{\mathbf{a}}(\tilde{\xi})$ は $O(\mathbf{a})$ のオーダー, $l_b(\tilde{\xi})$, $l_{\mathbf{v}}(\tilde{\xi})$, $l_w(\tilde{\xi})$ は $O(\mathbf{a}^2)$ のオーダーである．高次の項を無視することによって R_0 の近傍における学習方程式について次が成り立つ ([9]).

$$
\begin{aligned}
\dot{\mathbf{a}} &= 2(1-b^2) H(\mathbf{v}^*, w^*) \mathbf{a}, \\
\dot{b} &= -\frac{b(1-b^2)}{w^{*2}} \mathbf{a}^2 H(\mathbf{v}^*, w^*) - \frac{2b(b^2+1)}{w^{*2}} H(\mathbf{v}^*, w^*) \mathbf{a}^2.
\end{aligned}
$$

ここでエネルギー関数を $h(\mathbf{a}) := \frac{1}{2} \mathbf{a}^{\mathrm{T}} \mathbf{a}$ と定める.
R_0 の近傍では $h(\mathbf{a})$ に対して次が成り立つ.

$$
\dot{h} = \mathbf{a}^{\mathrm{T}} \dot{\mathbf{a}} = \frac{2 w^{*2} (b^2 - 1)}{b(b^2+3)} \dot{b}.
$$

さらに $R_0 \cap R_1$ の近傍では $h(\mathbf{a})$ に対して次が成り立つ ([9]).

$$
\dot{h} = \frac{w^{*2} (b^2 - 1)}{b(b^2+1)} \dot{b}.
$$

特異領域の近くの学習の軌道について次が成り立つ.

定理 16 ([9]) (特異領域の近くの学習方程式の軌道) エネルギー関数 $h(\mathbf{a})$ に対して，特異領域の近くの学習方程式の軌道は以下で表される．
R_0 の近傍では次の式で表される．

$$h(\mathbf{a}) = \frac{2w^{*2}}{3} \log \frac{(b^2+3)^2}{|b|} + C.$$

$R_0 \cap R_1$ の近傍では次の式で表される．

$$h(\mathbf{a}) = w^{*2} \log \left(|b| + \frac{1}{|b|} \right) + C.$$

6.1.2 Guo らによるダイナミクスの分類

Guo らにより，学習を始めるパラメータの初期値によって，特異領域の近くの学習のダイナミクスは次の 5 つの場合に分かれる [15]．分類について表 6.1 に示す．

表 6.1 ダイナミクスの分類

ダイナミクス	学習の様子
Fast convergence	特異領域を通過せずに真の分布に早く収束する．
Overlap singularity	特異領域である $\mathbf{a} = \mathbf{0}$ で停滞して真の分布まで到達しない場合． 学習の停滞 (プラトー) が起きる．$\mathbf{w_1} = \mathbf{w_2}$ になる．
Cross elimination singurarity	初期値に対して真の分布が特異領域をまたいで反対側にある場合． 特異領域である $b = \pm 1$ で学習の停滞が起きる．
Near elimination singularity	初期値に対して真の分布が特異領域と同じ側にある場合． 特異領域である $b = \pm 1$ に近づき学習の停滞が起きる．
Output weight 0	特異領域である $b = \pm 1$ で停滞して真の分布まで到達しない場合． 学習の停滞が起きる．w_3 又は w_4 が 0 になる．

6.2　各ダイナミクスの実行例

以下の結果は [27]，[28] に基づく．

$-3 \leq x \leq 3$ 上の入力 X，$\sigma = 0.05$ の雑音 Z に対して，訓練データを次で定める．

$$0.25\tanh(0.2x) + 0.25\tanh(0.4x) + Z$$

変数変換を行うと，$a = 0.2$, $b = 0$, $v = 0.3$, $w = 0.5$ となる．

真の分布を次で定める．

$$q(y|x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{|y - (0.25\tanh(0.2x) + 0.25\tanh(0.4x))\,|^2}{2\sigma^2}\right).$$

6.2.1　Overlap singularity 現象

学習モデルのパラメータの初期値を $a = 0.15$, $b = -1.8$, $v = 0.3$, $w = 0.5$ とする．

損失関数を対数尤度比関数として，ニューラルネットワークを 200 回学習させる．

```
results1[a_, b_] := NetTrain[trainingNet[a, b, v0, w0],
<|"Input" -> dataX, "Output" -> enc[dataY]|>,
"RoundWeightsHistories", "TrainedNet", "RoundLossList",
LossFunction -> "Loss", Method ->
"ADAM", "InitialLearningRate" -> 0.1,
BatchSize -> 30, MaxTrainingRounds -> 200]
```

パラメータ a, b の配列を作成し，学習回数と臨界直線に対する変化を図 6.1 に示す．

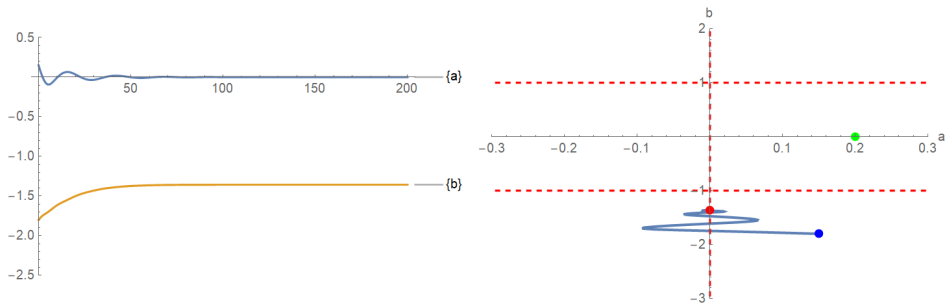


図 6.1　パラメータ a, b 変化

学習損失の配列を作成し，学習損失の変化と損失曲面上のダイナミクスを図 6.2 に示す．

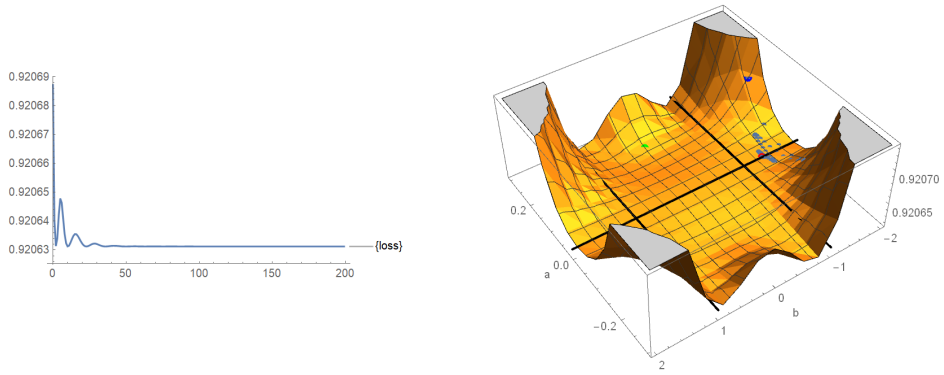


図 6.2　学習損失，学習損失曲面

6.2.2　Cross elimination singularity 現象

学習モデルのパラメータの初期値を $a = 0.15$, $b = -1.5$, $v = 0.3$, $w = 0.5$ とする．ニューラルネットワークを 100 回学習させる．

```
results1[a_, b_] := NetTrain[trainingNet[a, b, v0, w0],
<|"Input" -> dataX, "Output" -> enc[dataY]|>,
"RoundWeightsHistories", "TrainedNet", "RoundLossList",
LossFunction -> "Loss", Method ->
"ADAM", "InitialLearningRate" -> 0.1,
BatchSize -> 30, MaxTrainingRounds -> 100]
```

パラメータ a, b の配列を作成し，学習回数と臨界直線に対する変化を図 6.3 に示す．

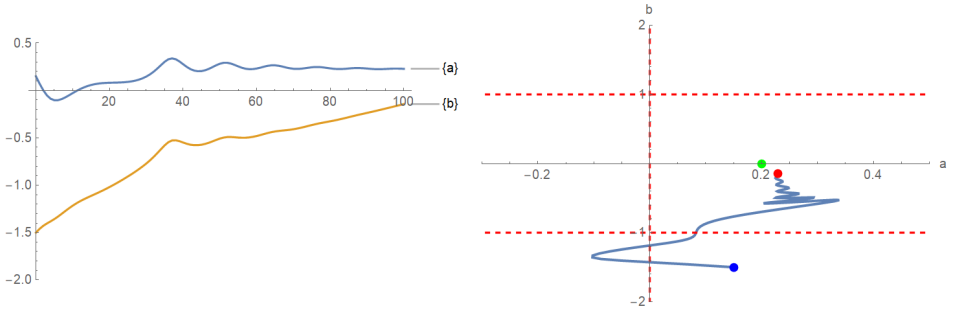


図 6.3　パラメータ a, b 変化

学習損失の配列を作成し，学習損失の変化と損失曲面上のダイナミクスを図 6.4 に示す．

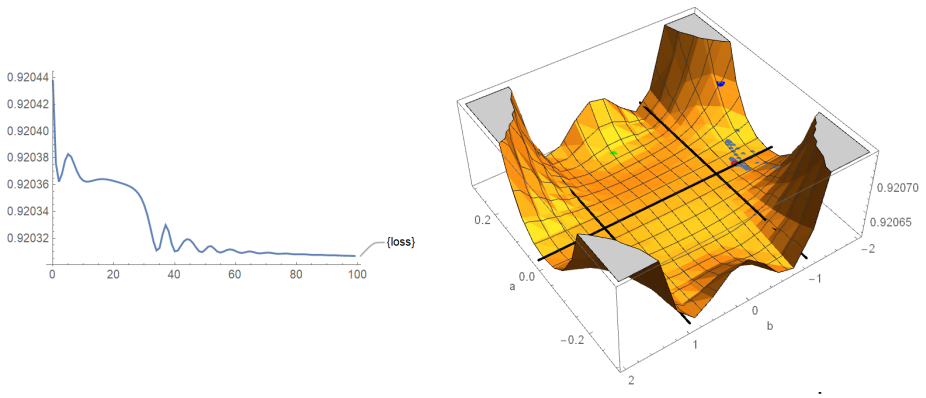


図 6.4　学習損失，学習損失曲面

6.2.3　Fast convergence 現象

学習モデルのパラメータの初期値を $a = 0.05,\ b = 0,\ v = 0.3,\ w = 0.5$ とする．ニューラルネットワークを 5 回学習させる．

```
results1[a_, b_] := NetTrain[trainingNet[a, b, v0, w0],
<|"Input" -> dataX, "Output" -> enc[dataY]|>,
"RoundWeightsHistories", "TrainedNet", "RoundLossList",
LossFunction -> "Loss", Method ->
"ADAM", "InitialLearningRate" -> 0. 1,
BatchSize -> 30, MaxTrainingRounds -> 5]
```

パラメータ $a,\ b$ の配列を作成し，学習回数と臨界直線に対する変化をグラフに表して図 6.5 に示す．

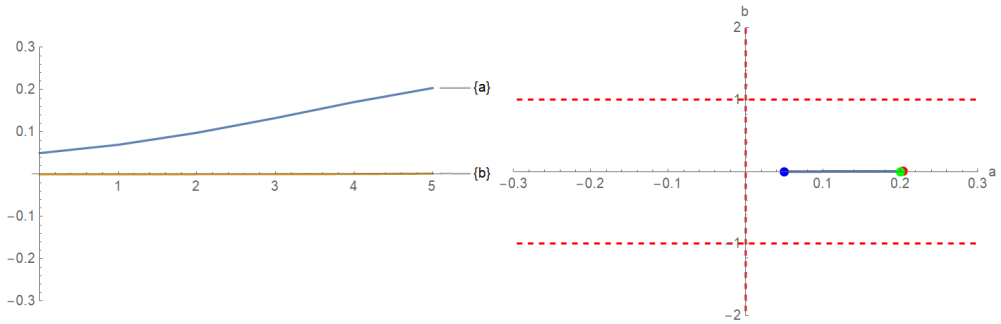


図 6.5　パラメータ $a,\ b$ 変化

学習損失の配列を作成し，学習損失の変化と損失曲面上のダイナミクスを図 6.6 に示す．

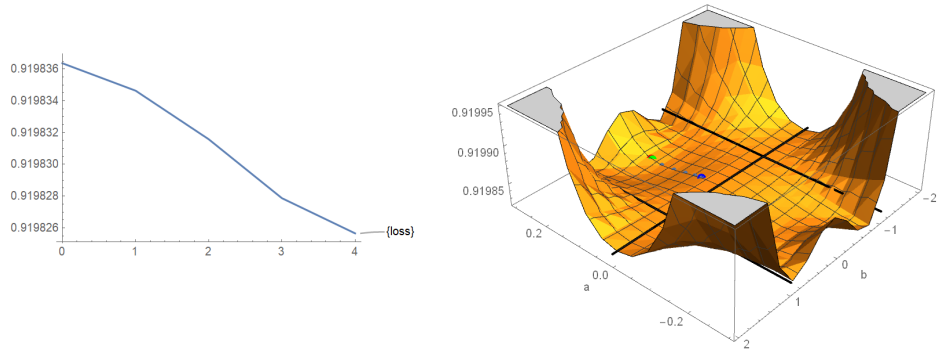


図 6.6　学習損失，学習損失曲面

6.2.4　Near elimination singularity 現象

学習モデルのパラメータの初期値を $a = 0.7$, $b = -0.6$, $v = 0.3$, $w = 0.5$ とする．ニューラルネットワークを 200 回学習させる．

```
results1[a_, b_] := NetTrain[trainingNet[a, b, v0, w0],
<|"Input" -> dataX, "Output" -> enc[dataY]|>,
"RoundWeightsHistories", "TrainedNet", "RoundLossList",
LossFunction -> "Loss", Method ->
"ADAM", "InitialLearningRate" -> 0. 1,
BatchSize -> 30, MaxTrainingRounds -> 200]
```

パラメータ a, b の配列を作成し，学習回数と臨界直線に対する変化を図 6.7 に示す．

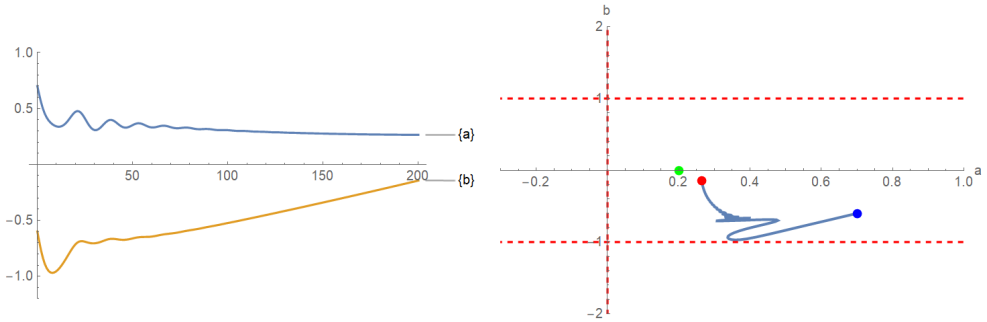


図 6.7　パラメータ a, b 変化

学習損失の配列を作成し，学習損失の変化と損失曲面上のダイナミクスを図 6.8 に示す．

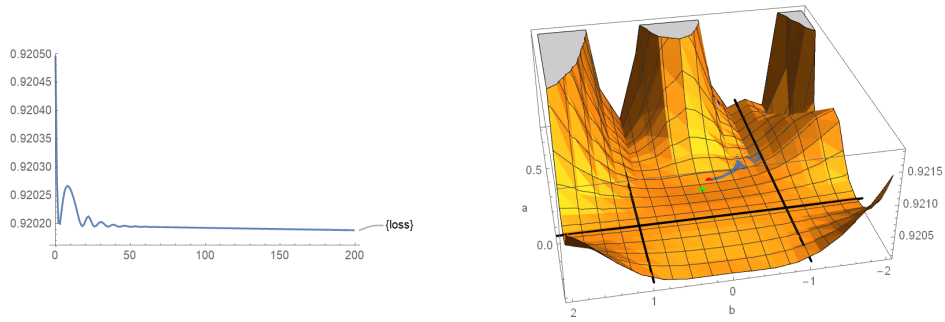


図 6.8　学習損失，学習損失曲面

6.2.5　Output weight 0 現象

学習モデルのパラメータの初期値を $a = 0.7,\ b = -0.5,\ v = 0.3,\ w = 0.5$ とする．ニューラルネットワークを 100 回学習させる．

```
results1[a_, b_] := NetTrain[trainingNet[a, b, v0, w0],
<|"Input" -> dataX, "Output" -> enc[dataY]|>,
"RoundWeightsHistories", "TrainedNet", "RoundLossList",
LossFunction -> "Loss", Method ->
"ADAM", "InitialLearningRate" -> 0. 1,
BatchSize -> 30, MaxTrainingRounds -> 100]
```

パラメータ a, b の配列を作成し，学習回数と臨界直線に対する変化を図 6.9 に示す．

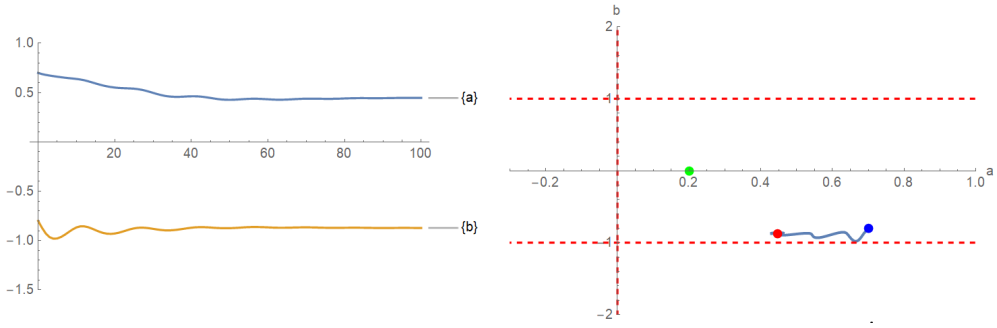


図 6.9　パラメータ a, b 変化

学習損失の配列を作成し，学習損失の変化と損失曲面上のダイナミクスを図 6.10 に示す．

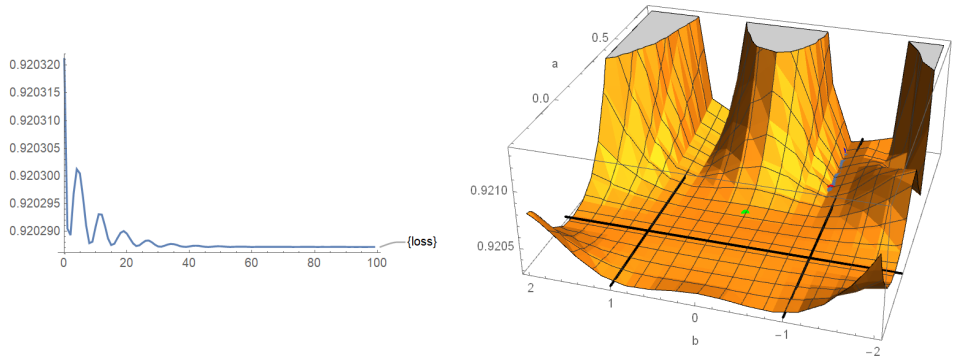


図 6.10　学習損失，学習損失曲面

6.3　シミュレーションによるダイナミクスの変化

Guo らによるダイナミクスの分類により, 初期値を変えるシミュレーションを行うことでダイナミクスが変化する様子を数学的に調べる.

学習モデルの初期値を特異領域に属する点の近傍より離れた所から特異領域 Overlap singularity と Elimination singularity に近づけることで，学習のダイナミクスの種類が変化することを調べる．以下の節は [28] に基づく.

6.3.1　例の設定

例 4 ([28])　(訓練データ, 真の分布)　$-3 \leq x \leq 3$ 上の入力 X，$\sigma = 0.05$ の雑音 Z に対して，訓練データを $0.25\tanh(0.2x) + 0.25\tanh(0.4x) + Z$ とする．真の分布の確率密度関数を次で定める.

$$q(y|x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{|y - (0.25\tanh(0.2x) + 0.25\tanh(0.4x))\,|^2}{2\sigma^2}\right).$$

訓練データの入力 x_s を次で定める:

$x_s = \{2.467685732795669, 1.6313896711002975, 1.7039693114471142, -2.353539095169551,$
$2.5106926463104458, -2.9742536653063, 1.4778884503387921, -1.7619315572659175,$
$0.8575206146347014, -1.9522402751318726, -2.9186556422433796, 2.3433244821789305,$
$-2.3174593747595598, 0.24745360478229195, -0.43473282858294837, 2.0777962243403962,$
$-0.7489587340884398, 0.40283200240701333, 1.4393667305075848, 2.6884952319559243,$

$0.4233060018195829, 1.3133371734415373, -1.8687861826912897, 2.641499809476027,$
$1.3536619131864676, 1.4261447937286373, -1.5373889449365947, 2.5833410435168336,$
$-0.7634883775841974, -1.418229957030034\}$,

訓練データの出力 y_s を次で定める:

$y_s = \{0.25047549494898375, 0.14195709642758433, 0.2416763776071971, -0.34055590890961035,$
$0.3082658314034902, -0.4292549244954509, 0.15038776701404105, -0.23410034295044008,$
$0.1674469014375939, -0.26548937037643955, -0.3321460817933551, 0.2720167181157782,$
$-0.2892455062624837, 0.0520546848151971, -0.0009290519327547, 0.2940081059525326,$
$-0.14421683321295234, 0.08562704853302514, 0.25724997978192643, 0.2668005655536598,$
$0.043918697553646746, 0.19753643159405437, -0.2627853499983649, 0.25989101875041354,$
$0.1395086144673041, 0.17062611740258554, -0.18466386529707135, 0.3690548490941195,$
$-0.16241114605952112, -0.14051890248769974\}$,

このとき訓練データを次で定める:

$\{2.46769 \to 0.250475, 1.63139 \to 0.141957, 1.70397 \to 0.241676, -2.35354 \to -0.340556,$
$2.51069 \to 0.308266, -2.97425 \to -0.429255, 1.47789 \to 0.150388, -1.76193 \to -0.2341,$
$0.857521 \to 0.167447, -1.95224 \to -0.265489, -2.91866 \to -0.332146, 2.34332 \to 0.272017,$
$-2.31746 \to -0.289246, 0.247454 \to 0.0520547, -0.434733 \to -0.000929052,$
$2.0778 \to 0.294008, -0.748959 \to -0.144217, 0.402832 \to 0.085627, 1.43937 \to 0.25725,$
$2.6885 \to 0.266801, 0.423306 \to 0.0439187, 1.31334 \to 0.197536, -1.86879 \to -0.262785,$
$2.6415 \to 0.259891, 1.35366 \to 0.139509, 1.42614 \to 0.170626, -1.53739 \to -0.184664,$
$2.58334 \to 0.369055, -0.763488 \to -0.162411, -1.41823 \to -0.140519\}$.

$a = 0.2$, $b = 0$, $v = 0.3$, $w = 0.5$ のとき，真の分布が学習モデルによって実現可能な場合，学習モデルの初期値を変えて，臨界直線の影響を受けて変化する損失関数を対数密度比関数として特異領域の近くの学習のダイナミクスを 5 つの場合に分類して考察する．

6.3.2 Overlap singularity 現象 と Cross elimination singularity 現象

学習モデルの初期値を $a = 0.15$, $b = -2.0, -1.8, -1.5, -1.3$, $v = 0.3$, $w = 0.5$ とし，ニューラルネットワークの学習を 140 回行う．臨界直線の影響を受け変化する a, b のパラメータの配列を作成する．a のパラメータの変化を図 6.11 の左側に，b のパラメータの変化を図 6.11 の中央に, a, b のパラメータの変化を図 6.11 の右側に示す.

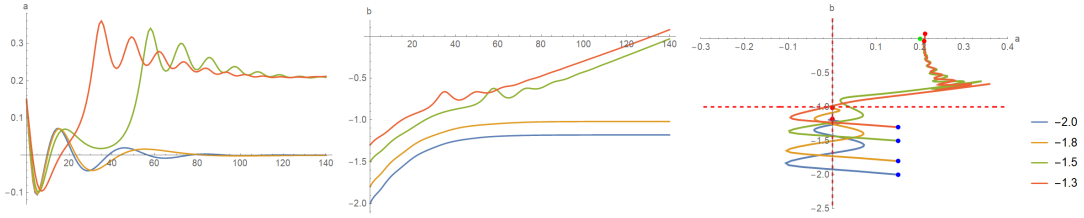


図 6.11　Evolution of parameters of a, b ($a = 0.15$, $b = -2.0, -1.8, -1.5, -1.3$)

ニューラルネットワークの学習を 140 回行う．学習損失の配列を作成し，学習損失を図 6.12 の左側に，学習損失曲面上の学習のダイナミクスを図 6.12 の右側に示す．

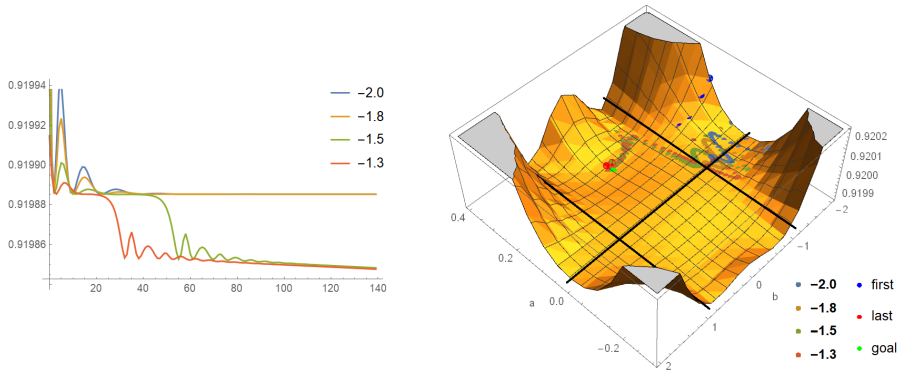


図 6.12　Evolution of the training loss and the dynamics of the training loss surface ($a = 0.15$, $b = -2.0, -1.8, -1.5, -1.3$)

パラメータ $b(-2.2 \leq b \leq 2.2)$ を動かしてシミュレーションする．a, b のパラメータの変化を図 6.13 の左側に，学習損失曲面上の学習のダイナミクスを図 6.13 の右側に示す．

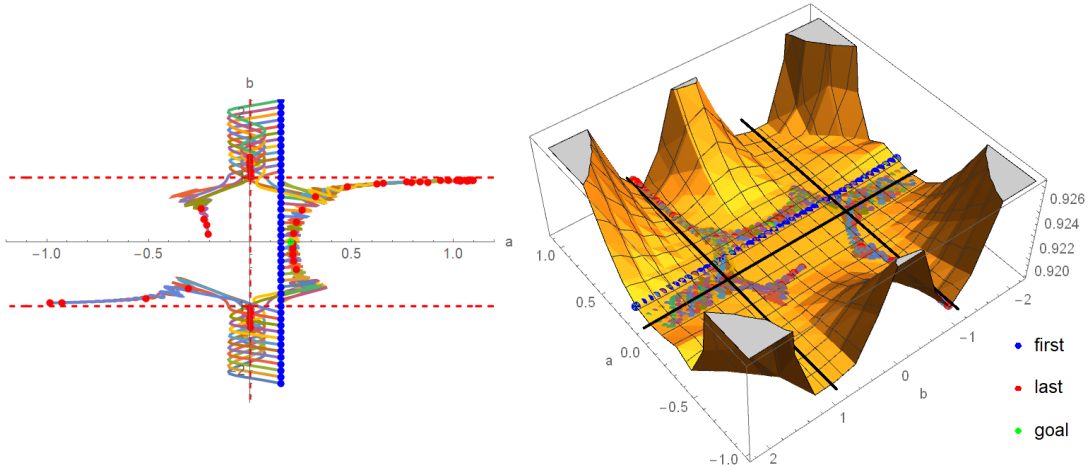


図 6.13　Evolution of the parameters of a, b and the dynamics of the training loss surface ($-2.2 \leq b \leq 2.2$)

命題 3 ([28]) (ダイナミクスの変化)
(1) 臨界直線 $a = 0$ 上でプラトー現象が起こり，$b = -2.0, -1.8$ の場合，学習のダイナミクスが真の分布に達しない．
(2) 臨界直線 $b = -1$ を越えるとプラトー現象が起こり，$b = -1.0, -1.8$ の場合，学習のダイナミクスが真の分布に達する．
(3) パラメータ b が 0 まで変化すると，学習のダイナミクスは Overlap singularity 現象から Cross elimination singularity 現象へ，Cross elimination singularity 現象から Fast convergence 現象へと変化する．

6.3.3 Near elimination singularity 現象 と Output weight 0 現象

学習モデルの初期値を $a = 0.5,\ 0.6,\ 0.7,\ 1.2, b = 0.75, v = 0.3, w = 0.5$ とする．ニューラルネットワークの学習を 100 回行う．臨界直線の影響を受けて変化する a, b のパラメータの配列を作成する．a のパラメータの変化を図 6.14 の左側に，b のパラメータの変化を図 6.14 の中央に表し，a, b のパラメータの変化を図 6.14 の右側に示す．

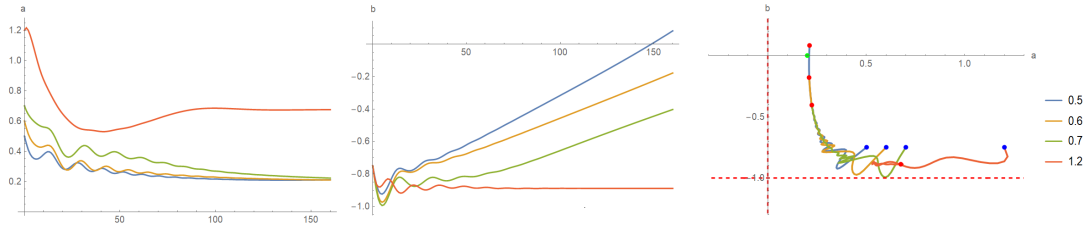


図 6.14　Evolutions of parameters of a, b ($a = 0.5,\ 0.6,\ 0.7,\ 1.2,\ b = 0.75$)

ニューラルネットワークの学習を 100 回行う．学習損失の配列を図 6.15 の左側に，学習損失曲面上の学習のダイナミクスを図 6.15 の右側に示す．

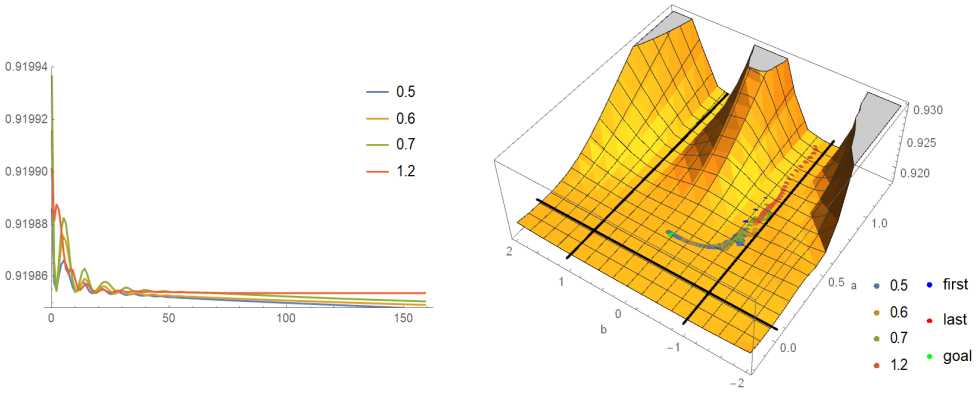


図 6.15　Evolution of the training loss and the dynamics of the training loss surface ($a = 0.5$, 0.6, 0.7, 1.2, $b = 0.75$)

パラメータ $a(0 \leq a \leq 2.2)$ を動かしてシミュレーションする．a, b のパラメータの変化を図 6.16 の左側に，学習損失曲面上の学習のダイナミクスを図 6.16 の右側に示す．

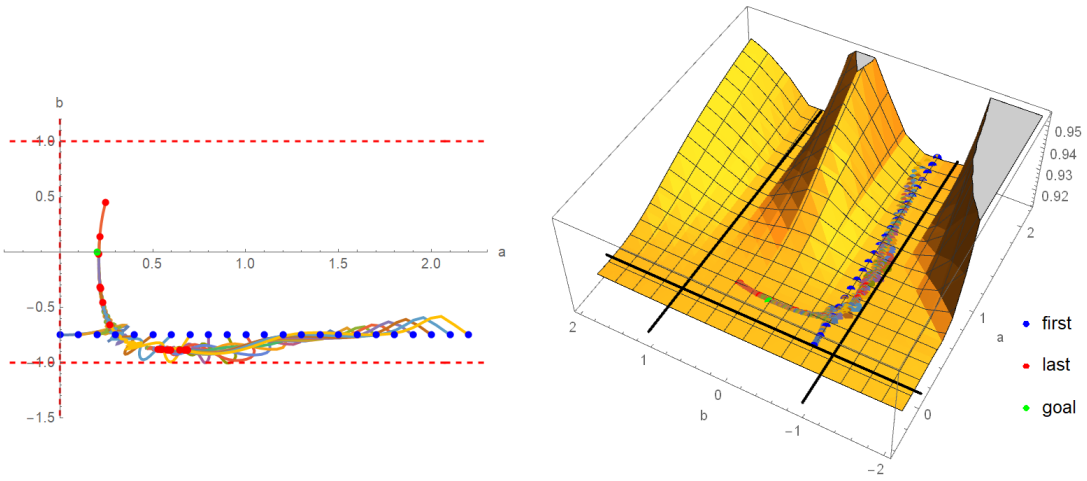


図 6.16　Evolution of the parameters of a, b and the dynamics of the training loss surface ($0 \leq a \leq 2.2$)

命題 4 ([28])　(ダイナミクスの変化)

(1) 臨界直線 $b = -1$ 上でプラトー現象が起こり，$a = 1.2$ の場合，学習のダイナミクスが真の分布に到達しない．

(2) 臨界直線 $b = -1$ に近づくとプラトー現象が起きて，$a = 0.6$, 0.7 では学習のダイナミクスが真の分布に達する．さらに，$a = 0.5$ の場合，学習のダイナミクスがより早く真の分布に到達する．

(3) a のパラメータが 0 になるにつれて，学習のダイナミクスは Output weight 0 現象から Near elimination singularity 現象へ，Near elimination singularity 現象から Fast convergence 現象へと変化する．

6.4　過学習・過剰般化が起きる場合の分析

第 2，3 章で示した過剰般化現象と過学習の現象の学習・汎化損失について調べる．
以下は [27] に基づく．

6.4.1　過学習を起こす学習のダイナミクス

例 5 ([27])　(訓練データ，テストデータ，真の分布)　$-3 \leq x \leq 3$ 上の入力 X，$\sigma = 0.05$ の雑音 Z に対して，訓練・テストデータを $0.25\tanh(3x) + 0.25\tanh(3x) + Z$ とする．真の分布を次で定める．

$$q(y|x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{|y\,(0.25\tanh(3x) + 0.25\tanh(3x))\,|^2}{2\sigma^2}\right).$$

訓練データを図 6.17 の左側に，テストデータを図 6.17 の右側に示す．

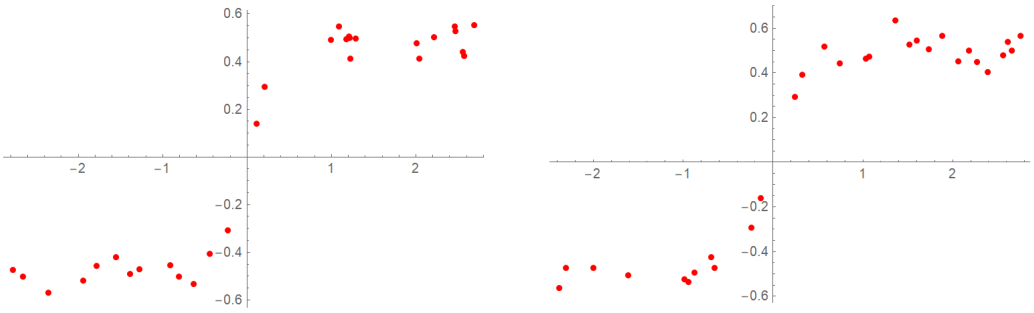


図 6.17　訓練データ，テストデータ

$a = 0$, $b = 0$, $c = 0$, $d = 0$, $v = 3$, $w = 0.5$ より真の分布が特異領域上にあるとき，過学習を起こす場合の学習のダイナミクスを考察する．
学習モデルの初期値を $a = 0.2$, $b = 0$, $c = 0$, $d = 0$, $v = 3$, $w = 0.5$ とすると，真の分布が学習モデルによって実現される．損失関数を 2 乗誤差関数，検証集合をテストデータとして，次のように入力してニューラルネットを 300 回学習させる．

```
results1[a_, b_] := NetTrain[trainingNet[a, b, 0, 0, v1, w1],
<|"Input" -> dataX, "Output" -> enc[dataY]|>, All, ValidationSet ->
<|"Input" -> testX, "Output" -> enc[testY]|>, LossFunction -> "Loss",
Method -> "ADAM", "InitialLearningRate" -> 0.1, BatchSize -> 30,
MaxTrainingRounds -> 300]
```

学習・汎化損失の配列を作成し，学習損失の変化を図 6.18 の左側に汎化損失の変化を図 6.18 の右側に示す．

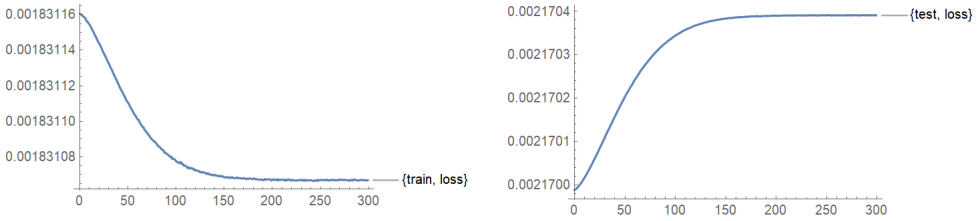


図 6.18　学習損失・汎化損失

学習損失が減少するが汎化損失が増加して，過学習が起きている．
学習・汎化損失の変化を図 6.19 の左側に，訓練・テストデータの上に学習後のニューラルネットの出力を図 6.19 の右側に示す．

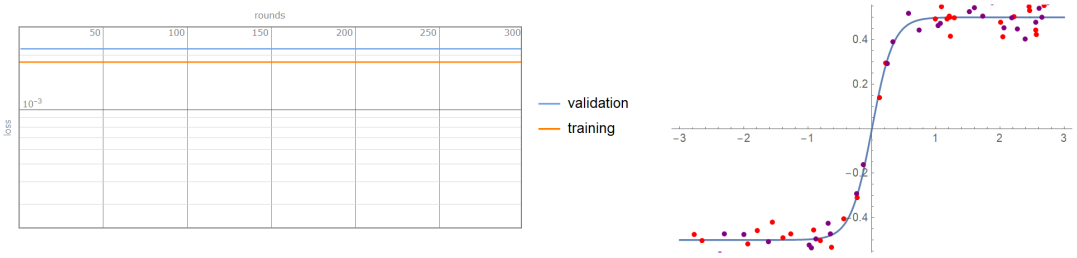


図 6.19　学習・汎化損失，学習後のニューラルネット

次のように入力してニューラルネットワークを 300 回学習させる.

```
results2[a_, b_] := NetTrain[trainingNet[a, b, 0, 0, v1, w1],
<|"Input"->dataX, "Output"->enc[dataY]|>, All, ValidationSet->
<|"Input"->testX, "Output"->enc[testY]|>, LossFunction->"Loss",
Method->"ADAM", "InitialLearningRate"->0.1, BatchSize->30,
MaxTrainingRounds->300]
```

パラメータ a, b の配列を作成し，学習回数に対する変化を図 6.20 の左側に臨界直線に対する変化を図 6.20 の右側に示す.

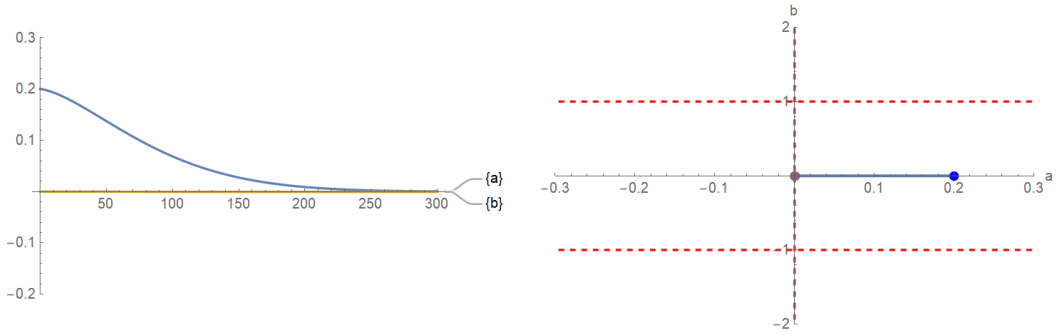


図 6.20　パラメータ a, b 変化

臨界直線 $a = 0$, $b = \pm 1$ に対して，学習損失曲面上のダイナミクスを図 6.21 の左側に汎化損失曲面上のダイナミクスを図 6.21 の右側に示す.

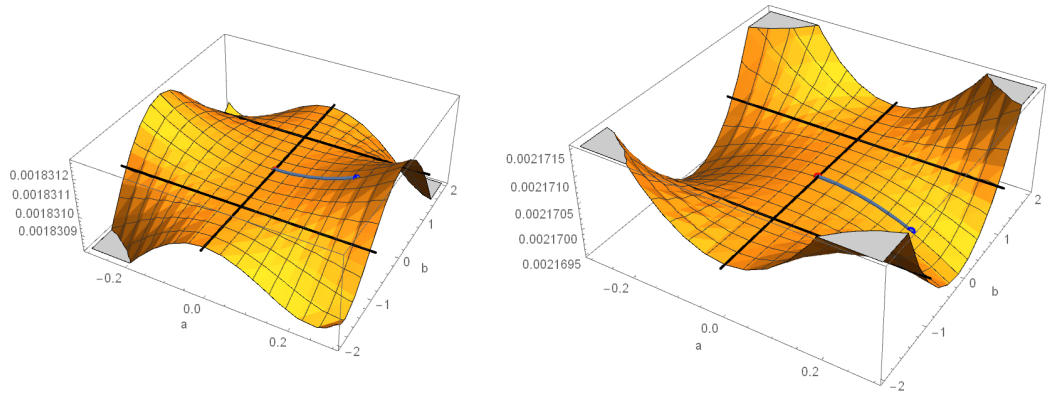


図 6.21　学習・汎化損失曲面上のダイナミクス

6.4.2　過剰般化を起こす学習のダイナミクス

学習が進むと関数を近似する際に過剰般化が起こる．その様子を図 6.22 に示す．

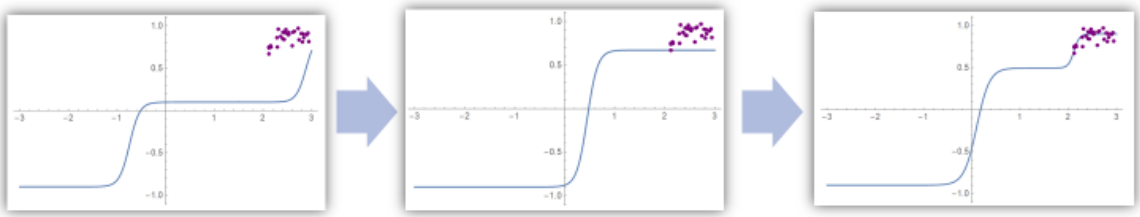


図 6.22　過剰般化の関数近似による分析

例 6 ([27])　[訓練データ, テストデータ, 真の分布] $-3 \leq x \leq 3$ 上の入力 X, $\sigma = 0.05$ の雑音 Z に対して, 訓練データを $0.5\tanh(4.8x)+0.4\tanh(5x-10)+Z$ と定め, $2.1 \leq x \leq 3$ 上の入力 X, 雑音 Z に対して, テストデータを $0.2\tanh(5x-10)+0.2\tanh(5x-10)+0.5+Z$ と定める．真の分布を次で定める．

$$q(y|x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{|y-(0.5\tanh(4.8x)+0.4\tanh(5x-10))|^2}{2\sigma^2}\right).$$

訓練データを図 6.23 の左側に，テストデータを図 6.23 の右側に示す．

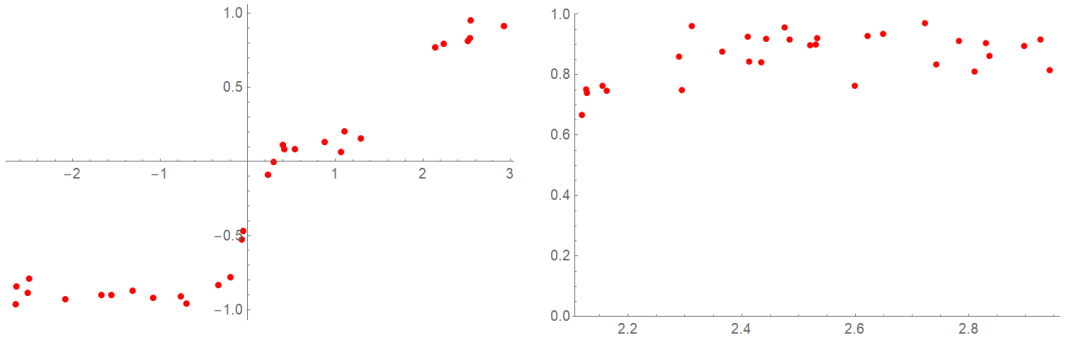


図 6.23　訓練データ, テストデータ

$a = 0.2$, $b = 0.11$, $c = -10$, $d = -4.44$, $v = 4.88$, $w = 0.9$ より真の分布が特異領域上の近くにあり．過剰般化を起こす場合の学習のダイナミクスを考察する．

学習モデルの初期値を $a = 0.2,\ b = 0.11,\ c = -18,\ d = -4.44,\ v = 4.88,\ w = 0.9$ とすると，$c = -18$ と固定するため，真の分布が学習モデルによって実現されない．損失関数を 2 乗誤差関数，検証集合をテストデータとして，次のように入力してニューラルネットを 20 回学習させる．

```
results1[a_, b_] := NetTrain[trainingNet[a, b, −18, −4.44, v1, w1],
< |"Input" − > dataX, "Output" − > enc[dataY]| >, All, ValidationSet− >
< |"Input" − > testX, "Output" − > enc[testY]| >, LossFunction− > "Loss",
Method− > "ADAM", "InitialLearningRate" − > 0.1, BatchSize− > 30,
MaxTrainingRounds− > 20]
```

学習・汎化損失の配列を作成し，学習損失の変化を図 6.24 の左側に汎化損失の変化を図 6.24 の右側に，学習・汎化損失の変化を図 6.25 に示す．

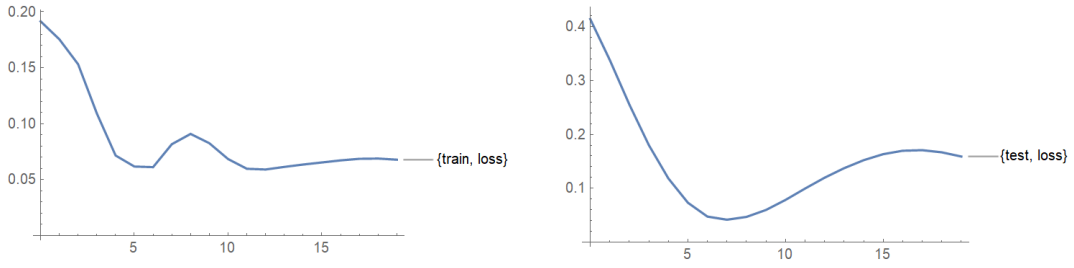


図 6.24　学習損失・汎化損失

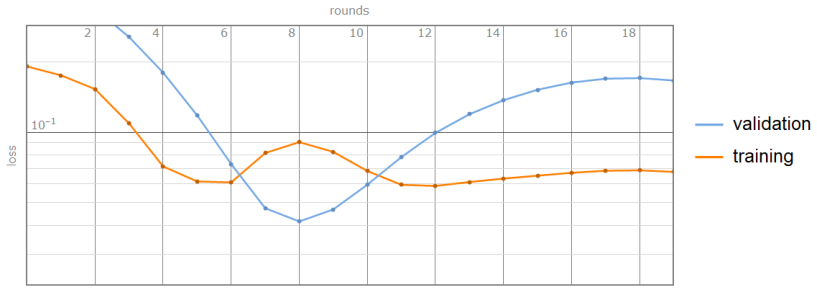


図 6.25　学習・汎化損失

汎化損失が減少して学習損失よりも小さくなり，増加して学習損失よりも大きくなる．テストデータの上にニューラルネットの出力について学習前を図 6.26 の左側に学習後を

図 6.26 の右側に示す.

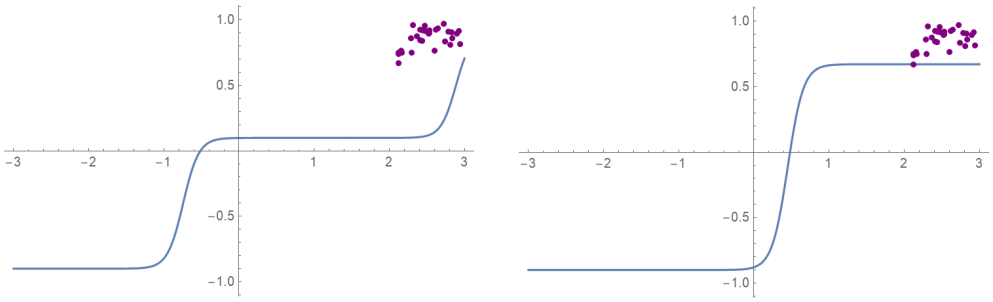


図 6.26　学習前・学習後のニューラルネット

次のように入力してニューラルネットワークを 70 回学習させる.

```
results2[a_, b_] := NetTrain[trainingNet[a, b, −18, −4.44, v1, w1],
<|"Input"− > dataX, "Output"− > enc[dataY]|>, All, ValidationSet− >
<|"Input"− > testX, "Output"− > enc[testY]|>, LossFunction− > "Loss",
Method− > "ADAM", "InitialLearningRate"− > 0.1, BatchSize− > 30,
MaxTrainingRounds− > 70]
```

パラメータ a, b の配列を作成し，学習回数に対する変化を図 6.27 の左側に臨界直線に対する変化を図 6.27 の右側に示す.

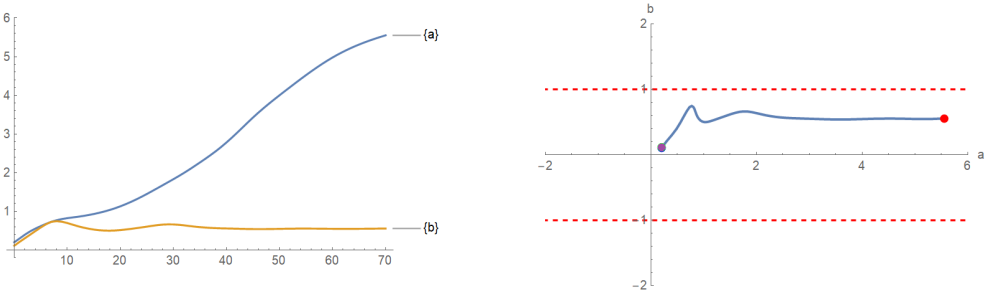


図 6.27　パラメータ a, b 変化

テストデータを検証集合として次のように入力してニューラルネットワークを 70 回学習させる.

```
results3[a_, b_] := NetTrain[trainingNet[a, b, −18, −4.44, v1, w1],
< |"Input" − > dataX, "Output" − > enc[dataY]| >, All, ValidationSet− >
< |"Input" − > testX, "Output" − > enc[testY]| >, LossFunction− > "Loss",
Method− > "ADAM", "InitialLearningRate" − > 0.1, BatchSize− > 30,
MaxTrainingRounds− > 70]
```

学習・汎化損失の変化を図 6.28 の左側に，テストデータの上に学習後のニューラルネットの出力を図 6.28 の右側に示す.

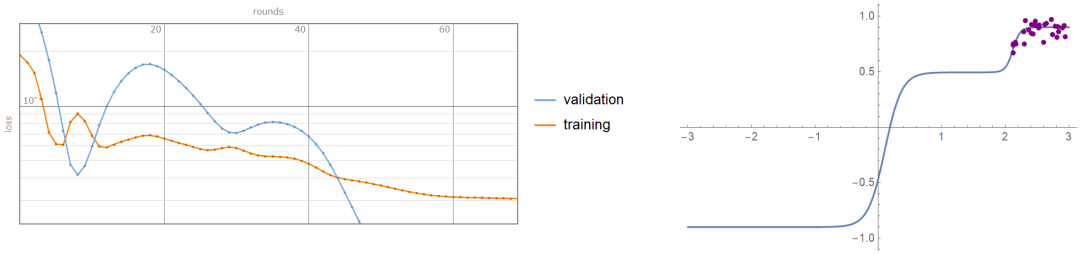


図 6.28　学習・汎化損失，学習後のニューラルネット

汎化損失が再び減少して小さくなる.
臨界直線を $b = \pm 1$ に対して，学習損失曲面上のダイナミクスを図 6.29 の左側に汎化損失曲面上のダイナミクスを図 6.29 の右側に示す.

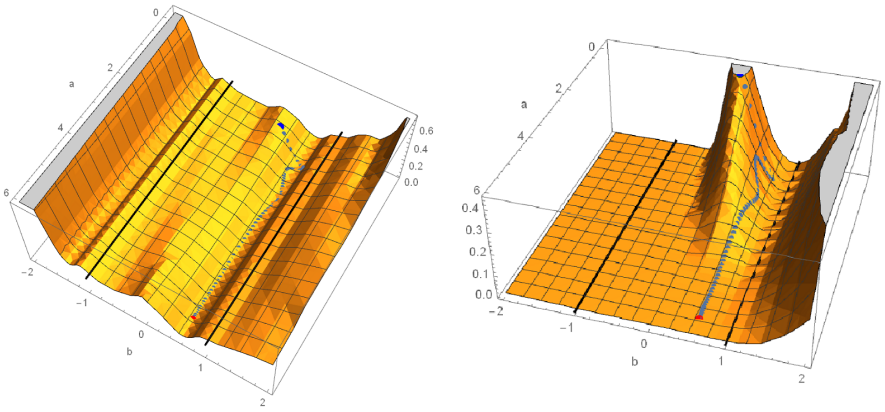


図 6.29　学習・汎化損失曲面上のダイナミクス

第 7 章

理解構造を捉える方法

中間ユニット数 $H = 2 \to H_0 = 1$ の場合について，数学の学習における過剰般化現象を学習損失曲面上に可視化する方法を考案する．甘利 ([16]) によれば，このような構造は一般の深層回路に到る所に埋め込まれている．

7.1　分析のための準備

「順列と組合せ」における記号計算，意味理解を問う考査問題を作成する．順列又は組合せの一方が満点で一方が部分点である生徒の得点をテストデータとして定め，ニューラルネットワークの重みを w_1, w_2, w_3, w_4 と定める．以下は副論文 [29]，副論文 [30] に基づく．

7.1.1　考査問題

数学 A の「場合の数と確率」の単元における「順列と組合せ」の概念理解について, 高校 1 年生対象に 3 回 (147 人, 152 人, 143 人) の定期考査のテストの結果により分析する．第 1 回考査は, 順列を学習した後, 組合せの学習も終えた段階で行う．第 2 回考査は順列や組合せを学習した後，確率を学習する段階 (組合せの学習後，時間が経過している) で行う． 第 3 回考査は確率を学習後時間が経過した時期に実施する．定期考査において, 以下の問題を与えた.

第1回定期考査

(1) 次の値を求めよ。

(ア) ${}_5\mathrm{P}_3$　(イ) $6!$　(ウ) ${}_6\mathrm{C}_4$　(エ) ${}_{20}\mathrm{C}_{18}$

第2回定期考査

(1) 次の値を求めよ。

(ア) ${}_5\mathrm{P}_3$　(イ) $6!$　(ウ) ${}_8\mathrm{C}_3$　(エ) ${}_{16}\mathrm{C}_{14}$

第3回定期考査

(1) 次の値を求めよ。

(ア) ${}_8\mathrm{P}_4$　(イ) ${}_7\mathrm{C}_3$

第1～3回定期考査

(2) 次のような並び方や選び方の総数を求めよ。

(ア) 9 人から 3 人を選んで 1 列に並べる。

(イ) 7 枚のカードの中から 2 枚を選ぶ。

(ウ) 4 人の生徒全員を 1 列に並べる。

(エ) 10 人の委員の中から委員長，副委員長，書記を 1 人ずつ選ぶ方法は，何通りあるか。ただし，兼任は認めないものとする。

7.1.2 意味理解と記号計算について

順列について，並べることを取り出すことに余り注意させずに導入する．1 つ 1 つのマス目に何通り入るかの積の法則として，全体の場合の数が求まる．そこで記号 ${}_n\mathrm{P}_r$ の導入と値の計算を行う．1 つの並び方も総数もイメージしやすい．その後役職を決める事や，色を塗り分け問題を通して「並べる」適応範囲を広げている．組合せについては取り出すことで順序を無視した組であると注意して導入する．1 つの選び方は具体的に考えやすいが総数はイメージしにくい．このとき，順列の総数を組合せを作り，並べたものの積の法則で考察している．並べることが，取り出した後で 1 組を並べることであると取り出すことに注意して考察する．この時互いの関係を考えて組合せの記号 ${}_n\mathrm{C}_r$ 導入して，総数を求める．このとき記号計算を数学の問題を分類する用語として考えたとき，順列と組合せの記号を正しく計算をすることができるかを判断する問題とする．記号の導入時に順列は並べる概念を用いて簡単に導入しているが，組合せは同値類で割る概念が難しいため深い理解が必要である．このとき意味理解を数学の問題を分類する用語として考えたとき，問題文にある「並べる」と「選ぶ」の言葉の意味を正しく考察できるかを判断する問題とする．「選ぶ」意味について，順序を無視したものであると順列の関係性を考える必要がある．「並べる」意味について，適応範囲を広げることから深い理解が必要である．

記号計算と意味理解において，組合せの学習がすぐに身に付くのか，過剰般化は起こりやすいのか，共に理解した後で理解が下がってしまうことはないかについて考える必要がある．

7.1.3　配点

生徒の正答を 1 次元ベクトルで表す．ここで, 記号計算を問 1 で考察する．順列の記号計算として問 (1)(ア), (イ), 組合せの記号計算として問 (1) (ウ), (エ) とする．次に, 意味理解を問 (2) で考察する．順列の意味理解として問 (2)(ア), (ウ), 組合せの意味理解として問 (2)(イ) とする．

順列と組合せのそれぞれの合計点が 1 となるように各成分の配点について, 記号計算と意味理解について順列の合計点 0.5 点，組合せの合計点 0.5 点とそれぞれ設定する．但し, 準正答は問 (1)(ア), 問 (2)(ア), (エ) を組合せ, 問 (1)(ウ), (エ), 問 (2)(イ) を順列，問 (2)(ウ) を 4! とせず 4 として計算したものとする．配点を次の図 7.1 のように設定する．

第1,2回考査

記号(順列)	正答	準正答
問1(ア)	0.4	0.2
問1(イ)	0.1	

記号(組合せ)	正答	準正答
問1(ウ)	0.3	0.15
問1(エ)	0.2	0.1

第3回考査

記号(順列)	正答	準正答
問1(ア)	0.5	0.25

記号(組合せ)	正答	準正答
問1(イ)	0.5	0.25

第1～3回考査

意味(順列)	正答	準正答
問2(ア)	0.4	0.2
問2(ウ)	0.1	0.05

意味(組合せ)	正答	準正答
問2(イ)	0.5	0.25

図 7.1　配点 (正答と準正答)

7.1.4　相互関係の問題について

相互関係の問題とは問 (2)(エ)「10 人の中から 3 つの役職を 1 人ずつ選ぶ総数を求める」とする．順列と組合せは異なる概念ではなく共通する概念を含んでいる．このとき相互関係を数学の問題を分類する用語として考えたとき，順列と組合せの共通の概念を考察して答えを導き出す必要のある問題となる．相互関係の問題について正答は順列で計算したもの，準正答は組合せで計算したものとする．

7.1.5　学習段階について

このとき意味理解について以下のように 4 つの学習段階を定める．
第 1 回考査から第 2 回考査にかけて以下の 2 つの段階を定める．

(1) 順列を学習後，組合せの学習が進まなかったが，相互関係の問題を正答した生徒
(2) 組合せの学習後，組合せの学習の影響を受けて相互関係の問題を準正答した生徒

第 2 回考査から第 3 回考査にかけて以下の 2 つの段階を定める．

(3) 過剰般化が起こり相互関係の問題を準正答した生徒 (組合せの学習が進む)
(4) 順列の再学習により組合せとの相互関係を考えて相互関係の問題を正答した生徒

このとき記号計算について以下のように 4 つの学習段階を定める．
第 1 回考査から第 2 回考査にかけて以下の 2 つの段階を定める．

(1) 順列を学習後，組合せの学習が進んで相互関係の問題を正答した生徒
(2) 過剰般化が起こりにくいが相互関係の問題を準正答した生徒 (組合せの学習が進まない)

第 2 回考査から第 3 回考査にかけて以下の 2 つの段階を定める．

(3) 過剰般化の影響を受けて相互関係の問題を準正答した生徒 (組合せの学習が進まない)
(4) 組合せの学習後，順列との相互関係を考えて相互関係の問題を正答した生徒

7.1.6　生徒集団の定義

テストデータを以下で定める．各考査で記号計算と意味理解において，生徒の解答について, 順列と組合せ分野の得点が共に 0.5 点 (満点) 未満の生徒を考察の対象から除く．よって, 次のように生徒集団 (i)，(ii)，(iii) を定める．

(i) 順列分野が満点で組合せ分野が部分点である生徒
(ii) 順列分野が部分点で組合せ分野が満点である生徒
(iii) 順列と組合せ分野が共に満点である生徒

(i)，(ii)，(iii) から 2 つの生徒集団を考察の対象とする．
4 つの学習段階 (1)〜(4) に対して，(1) は (i) と (iii)，(2) は (i) と (ii)，(3) は (i) と (ii)，(4) は (ii) と (iii) として，それぞれ 2 つの生徒集団を比較して考察する．

7.1.7　ニューラルネットワークの重みの定義

2 つの生徒集団のうち組合せの平均点を c 点，順列の平均点を d 点，生徒の人数を e 人，f 人とする．このとき，以下のようにニューラルネットワークの重みを定める．

$$w_1' = c,\ w_2' = d,\ w_3 = \frac{e}{e+f},\ w_4 = \frac{f}{e+f}.$$

次に部分集合として 3 回の各考査の生徒の状況を考察する．このとき第 1 回から第 3 回考査までの全考査と 3 回の各考査のパラメータ v, w を一定にするために次の補正を各考査の重み w_1', w_2' に作用させる．

ここで 2 つの生徒集団の生徒に対して第 1 回考査から第 3 回考査までの順列と組合せの合計の平均から 0.5 を引いた得点を g 点，各考査の順列と組合せの合計の平均から 0.5 を引いた得点をそれぞれ h_1, h_2, h_3 点とする．このとき以下の補正を定める．

$$w_i = w_i' \times \frac{g}{h_i}.$$

7.1.8　パラメータ a, b

初めにパラメータ a を，順列の平均点から組合せの平均点の差として定め，順列を組合せとする過剰般化と組合せを順列とする過剰般化のバランスについて考察する． パラメータ b を考察する 2 つの生徒集団の人数の割合と定める．順列が部分点で組合せが満点である生徒集団，順列が満点で組合せが部分点である生徒集団と，順列，組合せ共に完全正答した生徒集団から 2 つの集団を選んで人数の割合を考察する．a, b について次の図 7.2 に表す．

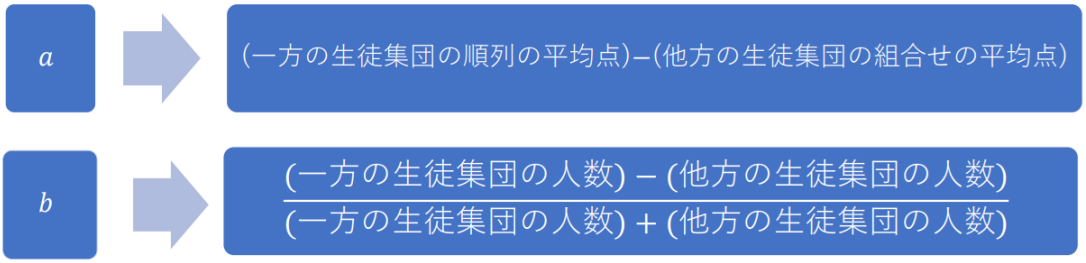


図 7.2　a, b の意味付け

2 つの生徒集団に対して a を「考査の平均点の差」とする．

考査の平均点 (順列と組合せの合計の平均点) は (順列の平均点) と (組合せの平均点) の和である．一方の生徒集団の組合せの平均点を w_1, 他方の生徒集団の順列の平均点を w_2 とする．生徒集団 (i), (ii), (iii) に対して平均点を次で表す．(但し, 順列と組合せの満点はそれぞれ 0.5 点とする．)

集団	対象	考査の平均点
(i)	順列 満点, 組合せ 部分点	0.5 + (組合せの平均点)
(ii)	順列 部分点, 組合せ 満点	(順列の平均点) + 0.5
(iii)	順列 満点, 組合せ 満点	0.5 + 0.5

次の (1)，(2)，(3) の場合に考査の平均点の差を考える．

(1) 集団 (i)，(ii) の場合:

$$\begin{aligned}&\text{「(ii) の平均点」} - \text{「(i) の平均点」}\\ &= \{(\text{順列の平均点}) + 0.5\} - \{0.5 + (\text{組合せの平均点})\}\\ &= (\text{順列の平均点}) - (\text{組合せの平均点})\end{aligned}$$

(2) 集団 (ii)，(iii) の場合:

$$\begin{aligned}\text{「(ii) の平均点」} - \text{「(iii) の平均点」} &= \{0.5 + (\text{順列の平均点})\} - (0.5 + 0.5)\\ &= (\text{順列の平均点}) - 0.5\\ &= (\text{順列の平均点}) - (\text{組合せの平均点})\end{aligned}$$

(3) 集団 (i)，(iii) の場合:

$$\begin{aligned}\text{「(iii) の平均点」} - \text{「(i) の平均点」} &= (0.5 + 0.5) - \{0.5 + (\text{組合せの平均点})\}\\ &= 0.5 - (\text{組合せの平均点})\\ &= (\text{順列の平均点}) - (\text{組合せの平均点})\end{aligned}$$

よって，3 集団のうちの 2 つの生徒集団に対して，「考査の平均点の差」と「一方の集団の順列の平均点と他方の集団の組合せの平均点の差」は等しい．

よって $a = w_2 - w_1$ を「2 つの生徒集団の考査の平均点の差」とする．

7.2　考査の結果と学習損失曲面上への可視化

2 つの概念 (順列と組合せ) に対して，生徒集団の理解の状況を情報科学における学習損失曲面上に可視化する．

7.2.1　学習損失曲面の描画方法

生徒集団と考査に対してニューラルネットワークの生徒の合計点 (入力) と，順列と組合せの平均点, 生徒の割合から求めた重みから出力を定める．真の分布と a, b をパラメータとする学習モデルに対して，対数尤度比関数を損失関数としてカルバック情報量の値を曲面上に表し，その曲面を学習損失曲面と定める．真の分布との差 (理解度の差) を損失として高さで表すと，真の分布の学習状況に近づけば小さい値を取り，遠ざかれば大きな値をとる．学習損失曲面 ($-1 \leq b \leq 1$) の性質として以下の左右非対称性を持っている．順列をすべて正答する生徒が増えれば (b が増加すれば) 組合せを順列とする過剰般化が増える (a は減少する)，また組合せをすべて正答する生徒が増えれば (b が減少すれば) 順列を組合せとする過剰般化が増える (a は増加する).

7.2.2　考査の結果 (意味理解)

意味理解について相互関係の問題を正答 (左側 4 列)，準正答 (右側 2 列) した生徒の平均点, 人数, 補正係数について以下の図 7.3 のように設定する．

第1回〜第3回	(i)(正答,部分点)	(iii)(正答,正答)	(ii)(部分点,正答)	(iii)(正答,正答)	(i)(正答,部分点)	(ii)(部分点,正答)
平均 c,d	0.228571429	0.5	0.26	0.5	0.22826087	0.279411765
人数 e,f	70	57	20	57	23	85
全体平均 g	0.350393701		0.437662338		0.268518519	
全体人数	127		77		108	

第1回	(i)(正答,部分点)	(iii)(正答,正答)	(ii)(部分点,正答)	(iii)(正答,正答)	(i)(正答,部分点)	(ii)(部分点,正答)
平均 c,d	0.234126984	0.5	0.257142857	0.5	0.214285714	0.245
人数 e,f	63	5	7	5	7	10
全体平均 h_1	0.253676471		0.358333333		0.232352941	
全体人数	68		12		17	
補正係数 $\frac{g}{h_1}$	1.381262125		1.221383268		1.15564932	
補正 w_1, w_2	0.323390735	0.690631062	0.314069983	0.610691634	0.24763914	0.283134083

第2回	(i)(正答,部分点)	(iii)(正答,正答)	(ii)(部分点,正答)	(iii)(正答,正答)	(i)(正答,部分点)	(ii)(部分点,正答)
平均 c,d	0.25	0.5	0.283333333	0.5	0.25	0.276666667
人数 e,f	3	23	6	23	8	60
全体平均 h_2	0.471153846		0.455172414		0.273529412	
全体人数	26		29		68	
補正係数 $\frac{g}{h_2}$	0.743692753		0.961530893		0.981680605	
補正 w_1, w_2	0.185923188	0.371846376	0.272433753	0.480765447	0.245420151	0.271598301

第3回	(i)(正答,部分点)	(iii)(正答,正答)	(ii)(部分点,正答)	(iii)(正答,正答)	(i)(正答,部分点)	(ii)(部分点,正答)
平均 c,d	0.125	0.5	0.242857143	0.5	0.21875	0.313333333
人数 e,f	4	29	7	29	8	15
全体平均 h_3	0.454545455		0.45		0.280434783	
全体人数	33		36		23	
補正係数 $\frac{g}{h_3}$	0.770866142		0.972582973		0.957507895	
補正 w_1, w_2	0.096358268	0.385433071	0.236198722	0.486291486	0.209454852	0.30001914

図 7.3　意味理解

7.2.3　集団 (i) と (iii) の入力 (意味理解)

各生徒の順列と組合せの合計点を入力として，第 1 回考査から第 3 回考査までの入力 x_1，第 1 回考査の入力 x_2，第 2 回考査の入力 x_3，第 3 回考査の入力 x_4 を次で定める．

$$x_1 := \{1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.5,0.5,0.5,0.5,0.75,0.75,0.5,0.5,0.75,0.75,0.75\}$$

$$x_2 := \{1,1,1,1,1,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.75,0.5,0.5,0.5,0.5\}$$

$$x_3 := \{1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0.75,0.75,0.75\}$$

$x_4 := \{1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0.75,0.75,0.5,0.5\}$

3 層ニューラルネットワークの出力を次で定める.

$$y = w_3 tanh(w_1 x_i) + w_4 tanh(w_2 x_i).$$

また第 1 回考査について重みを次で定める.

$$w_1 = 0.323390735,\ w_3 = \frac{63.0}{68};\ w_2 = 0.690631062,\ w_4 = \frac{5.0}{68};$$

このとき $a = 0.36724,\ b = 0.852941$ である.

また第 2 回考査について重みを次で定める.

$$w_1 = 0.185923188,\ w_3 = \frac{3.0}{26};\ w_2 = 0.371846376,\ w_4 = \frac{23.0}{26};$$

このとき $a = 0.185923,\ b = -0.769231$ である.

また第 3 回考査について重みを次で定める.

$$w_1 = 0.096358268,\ w_3 = \frac{4.0}{33};\ w_2 = 0.385433071,\ w_4 = \frac{29.0}{33};$$

このとき $a = 0.289075,\ b = -0.757576$ である.

このとき，パラメータ v, w を $v_0 := 0.350394,\ w_0 := 1$ で固定させ，パラメータ a, b の初期値を第 1 回から第 3 回で定める.

7.2.4 集団 (i) と (ii) の入力 (意味理解)

各生徒の順列と組合せの合計点を入力として，第 1 回考査から第 3 回考査までの入力 x_1，第 1 回考査の入力 x_2，第 2 回考査の入力 x_3，第 3 回考査の入力 x_4 を次で定める.

$x_1 := \{0.6, 0.8, 0.8, 0.8, 0.95, 0.5, 0.7, 0.7, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.75, 0.5, 0.8, 0.75, 0.75, 0.8, 0.8, 0.6, 0.8, 0.8, 0.8, 0.8, 0.8, 0.6, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.75, 0.8, 0.8, 0.8, 0.7, 0.6, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.9, 0.6, 0.8, 0.9, 0.8, 0.8, 0.95, 0.8, 0.75, 0.75, 0.75, 0.75, 0.75, 0.75, 0.75, 0.75, 0.75, 0.5, 0.75, 0.75, 0.75, 0.75, 0.75, 0.75, 0.75, 0.75, 0.75, 0.8, 0.8, 0.8, 0.6, 0.8, 0.6, 0.8, 0.75, 0.7, 0.9, 0.8, 0.8, 0.75, 0.6, 0.8, 0.8, 0.8, 0.95, 0.8, 0.7\}$

$x_2 := \{0.6, 0.8, 0.8, 0.75, 0.5, 0.75, 0.8, 0.75, 0.75, 0.75, 0.95, 0.5, 0.7, 0.75, 0.7, 0.8, 0.8\}$

$x_3 := \{0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.6, 0.8, 0.8, 0.8, 0.8, 0.8, 0.6, 0.8, 0.8, 0.8, 0.8, 0.75, 0.75, 0.8, 0.8, 0.8, 0.75, 0.8, 0.8, 0.8, 0.7, 0.75, 0.6, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.75, 0.8, 0.75, 0.8, 0.75, 0.75, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.6, 0.8, 0.6, 0.8, 0.7, 0.9, 0.8, 0.8, 0.75, 0.6, 0.8, 0.75\}$

$x_4 := \{0.8, 0.75, 0.5, 0.75, 0.75, 0.75, 0.75, 0.95, 0.9, 0.9, 0.75, 0.8, 0.8, 0.8, 0.8, 0.8, 0.75, 0.6, 0.95, 0.8, 0.8, 0.8, 0.7\}$

3 層ニューラルネットワークの出力を次で定める.

$$y = w_3 tanh(w_1 x_i) + w_4 tanh(w_2 x_i).$$

また第 1 回考査について重みを次で定める.

$$w_1 = 0.24763914,\ w_3 = \frac{7.0}{17};\ w_2 = 0.283134083,\ w_4 = \frac{10.0}{17};$$

このとき $a = 0.0354949,\ b = -0.176471$ である.

また第 2 回考査について重みを次で定める.

$$w_1 = 0.245420151,\ w_3 = \frac{8.0}{68};\ w_2 = 0.271598301,\ w_4 = \frac{60.0}{68};$$

このとき $a = 0.0261781,\ b = -0.764706$ である.

また第 3 回考査について重みを次で定める.

$$w_1 = 0.209454852,\ w_3 = \frac{8.0}{23};\ w_2 = 0.30001914,\ w_4 = \frac{15.0}{23};$$

このとき $a = 0.0905643,\ b = -0.304348$ である.

このとき，パラメータ $v,\ w$ を $v_0 := 0.268519,\ w_0 := 1$ で固定させ，パラメータ $a,\ b$ の初期値を第 1 回から第 3 回で定める.

7.2.5 集団 (ii) と (iii) の入力 (意味理解)

各生徒の順列と組合せの合計点を入力として，第 1 回考査から第 3 回考査までの入力 x_1，第 1 回考査の入力 x_2，第 2 回考査の入力 x_3，第 3 回考査の入力 x_4 を次で定める.

$x_1 := \{1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0.8, 0.8, 0.8, 0.9, 0.8, 0.8, 0.8, 0.8, 0.8, 0.6, 0.5, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.7, 0.7, 0.6\}$

$x_2 := \{1, 1, 1, 1, 1, 0.8, 0.8, 0.8, 0.8, 0.8, 0.7, 0.6\}$

$x_3 := \{1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0.8, 0.8, 0.8, 0.8, 0.8, 0.7\}$

$x_4 := \{1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0.9, 0.8, 0.8, 0.8, 0.8, 0.6, 0.5\}$

3 層ニューラルネットワークの出力を次で定める.

$$y = w_3 tanh(w_1 x_i) + w_4 tanh(w_2 x_i).$$

また第 1 回考査について重みを次で定める.

$$w_1 = 0.610691634,\ w_3 = \frac{5.0}{12};\ w_2 = 0.314069983,\ w_4 = \frac{7.0}{12};$$

このとき $a = -0.296622,\ b = -0.166667$ である.

また第 2 回考査について重みを次で定める.

$$w_1 = 0.480765447,\ w_3 = \frac{23.0}{29};\ w_2 = 0.272433753,\ w_4 = \frac{6.0}{29};$$

このとき $a = -0.208332,\ b = 0.586207$ である.

また第 3 回考査について重みを次で定める.

$$w_1 = 0.486291486,\ w_3 = \frac{29.0}{36};\ w_2 = 0.236198722,\ w_4 = \frac{7.0}{36};$$

このとき $a = -0.250093,\ b = 0.611111$ である.

このとき，パラメータ v, w を $v_0 := 0.437662$, $w_0 := 1$ で固定させ，パラメータ a, b の初期値を第 1 回から第 3 回で定める.

7.2.6　学習損失曲面上への可視化 (意味理解)

ここで理解の状況を第 1 回考査 (青)，第 2 回考査 (赤)，第 3 回考査 (緑)，第 1 回から第 3 回までを (黄) で点として表示する (図 7.4)．　意味理解において生徒集団 (i),(iii) については図 7.4 の上側に (i),(ii) については中央に (ii),(iii) については下側に示す．

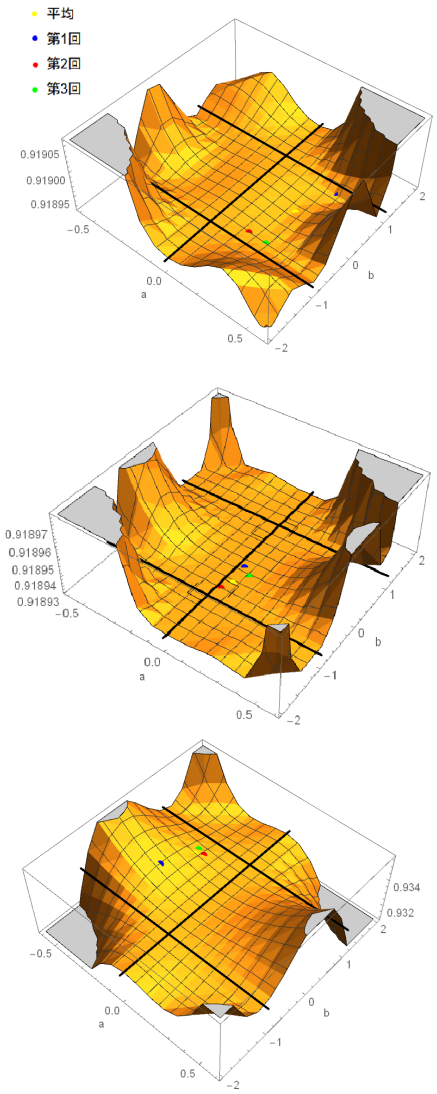


図 7.4　意味理解の学習損失曲面

7.2.7　考査の結果 (記号計算)

記号計算について相互関係の問題を正答 (左側 4 列)，準正答 (右側 2 列) した生徒の平均, 人数, 補正係数について以下の図 7.5 のように設定する．

第1回～第3回	(i)(正答,部分点)	(iii)(正答,正答)	(ii)(部分点,正答)	(iii)(正答,正答)	(i)(正答,部分点)	(ii)(部分点,正答)
平均 c,d	0.282352941	0.5	0.127777778	0.5	0.291071429	0.172727273
人数 e,f	34	121	9	121	28	11
全体平均 g	0.452258065		0.474230769		0.257692308	
全体人数	155		130		39	

第1回	(i)(正答,部分点)	(iii)(正答,正答)	(ii)(部分点,正答)	(iii)(正答,正答)	(i)(正答,部分点)	(ii)(部分点,正答)
平均 c,d	0.282142857	0.5	0.15	0.5	0.3	0.425
人数 e,f	28	50	6	50	7	2
全体平均 h_1	0.421794872		0.4625		0.327777778	
全体人数	78		56		9	
補正係数 $\frac{g}{h_1}$	1.072222767		1.025363825		0.786179922	
補正 w_1,w_2	0.302519995	0.536111383	0.153804574	0.512681913	0.235853977	0.334126467

第2回	(i)(正答,部分点)	(iii)(正答,正答)	(ii)(部分点,正答)	(iii)(正答,正答)	(i)(正答,部分点)	(ii)(部分点,正答)
平均 c,d	0.283333333	0.5	0	0.5	0.305263158	0.4
人数 e,f	6	28	0	28	19	2
全体平均 h_2	0.461764706		0.5		0.314285714	
全体人数	34		28		21	
補正係数 $\frac{g}{h_2}$	0.979412369		0.948461538		0.81993007	
補正 w_1,w_2	0.277500171	0.489706185	0	0.474230769	0.250294442	0.327972028

第3回	(i)(正答,部分点)	(iii)(正答,正答)	(ii)(部分点,正答)	(iii)(正答,正答)	(i)(正答,部分点)	(ii)(部分点,正答)
平均 c,d	0	0.5	0.083333333	0.5	0.125	0.035714286
人数 e,f	0	43	3	43	2	7
全体平均 h_3	0.5		0.472826087		0.055555556	
全体人数	43		46		9	
補正係数 $\frac{g}{h_3}$	0.904516129		1.002970822		4.638461538	
補正 w_1,w_2	0	0.452258065	0.083580902	0.501485411	0.579807692	0.165659342

図 7.5　記号計算

7.2.8　集団 (i) と (iii) の入力 (記号計算)

各生徒の順列と組合せの合計点を入力として，第 1 回考査から第 3 回考査までの入力 x_1，第 1 回考査の入力 x_2，第 2 回考査の入力 x_3，第 3 回考査の入力 x_4 を次で定める．

$x_1 := \{1, 0.7, 0.7, 0.6, 0.9, 0.8, 0.85, 0.8, 0.75, 0.8, 0.8, 0.8, 0.9, 0.65, 0.75, 0.9, 0.9, 0.9, 0.6, 0.7, 0.85, 0.75, 0.5, 0.9, 0.75, 0.85, 0.7, 0.8, 0.8, 0.75, 0.9, 0.9, 0.75, 0.8, 0.8\}$

$x_2 := \{1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0.7, 0.7, 0.85, 0.8, 0.75, 0.8, 0.8, 0.8, 0.9, 0.65, 0.75, 0.9, 0.9, 0.9, 0.6, 0.7, 0.85, 0.75, 0.5, 0.9, 0.75, 0.85, 0.7, 0.75, 0.9, 0.9, 0.75, 0.8\}$

$x_3 := \{1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0.6, 0.9, 0.8, 0.8, 0.8, 0.8\}$

$x_4 := \{1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1\}$

3 層ニューラルネットワークの出力を次で定める.

$$y = w_3 tanh(w_1 x_i) + w_4 tanh(w_2 x_i).$$

また第 1 回考査について重みを次で定める.

$$w_1 = 0.302519995,\ w_3 = \frac{28.0}{78};\ w_2 = 0.536111383,\ w_4 = \frac{50.0}{78};$$

このとき $a = 0.233591,\ b = -0.282051$ である.

また第 2 回考査について重みを次で定める.

$$w_1 = 0.277500171,\ w_3 = \frac{6.0}{34};\ w_2 = 0.489706185,\ w_4 = \frac{28.0}{34};$$

このとき $a = 0.212206,\ b = -0.647059$ である.

また第 3 回考査について重みを次で定める.

$$w_1 = 0.0,\ w_3 = \frac{0.0}{43};\ w_2 = 0.452258065,\ w_4 = \frac{43.0}{43};$$

このとき $a = 0.452258,\ b = -1$ である.

このとき，パラメータ v, w を $v_0 := 0.452258,\ w_0 := 1$ で固定させ，パラメータ a, b の初期値を第 1 回から第 3 回で定める.

7.2.9 集団 (i) と (ii) の入力 (記号計算)

各生徒の順列と組合せの合計点を入力として，第 1 回考査から第 3 回考査までの入力 x_1，第 1 回考査の入力 x_2，第 2 回考査の入力 x_3，第 3 回考査の入力 x_4 を次で定める.

$x_1 := \{0.8, 0.95, 0.9, 0.9, 0.9, 0.8, 0.8, 0.9, 0.5, 0.8, 0.8, 0.8, 0.5, 0.9, 0.8, 0.9, 0.7, 0.9, 0.9, 0.9, 0.9, 0.8, 0.8, 0.9, 0.9, 0.7, 0.8, 0.8, 0.7, 0.9, 0.5, 0.5, 0.75, 0.5, 0.5, 0.5, 0.5, 0.5, 0.75\}$

$x_2 := \{0.8, 0.95, 0.9, 0.9, 0.9, 0.9, 0.5, 0.8, 0.8\}$

$x_3 := \{0.7, 0.8, 0.9, 0.8, 0.5, 0.9, 0.9, 0.8, 0.7, 0.7, 0.9, 0.8, 0.9, 0.9,\ 0.8, 0.9, 0.9, 0.8, 0.8, 0.9, 0.8\}$

$x_4 := \{0.5, 0.5, 0.75, 0.5, 0.5, 0.5, 0.5, 0.5, 0.75\}$

3 層ニューラルネットワークの出力を次で定める.

$$y = w_3 tanh(w_1 x_i) + w_4 tanh(w_2 x_i).$$

また第 1 回考査について重みを次で定める.

$$w_1 = 0.235853977,\ w_3 = \frac{7.0}{9};\ w_2 = 0.334126467,\ w_4 = \frac{2.0}{9};$$

このとき $a = 0.0982725,\ b = 0.555556$ である.

また第 2 回考査について重みを次で定める.

$$w_1 = 0.250294442,\ w_3 = \frac{19.0}{21};\ w_2 = 0.327972028,\ w_4 = \frac{2.0}{21};$$

このとき $a = 0.0776776,\ b = 0.809524$ である.

また第 3 回考査について重みを次で定める.

$$w_1 = 0.579807692,\ w_3 = \frac{2.0}{9};\ w_2 = 0.165659342,\ w_4 = \frac{7.0}{9};$$

このとき $a = -0.414148,\ b = -0.555556$ である.

このとき，パラメータ v, w を $v_0 := 0.257692$, $w_0 := 1$ で固定させ，パラメータ a, b の初期値を第 1 回から第 3 回で定める.

7.2.10 集団 (ii) と (iii) の入力 (記号計算)

各生徒の順列と組合せの合計点を入力として，第 1 回考査から第 3 回考査までの入力 x_1，第 1 回考査の入力 x_2，第 2 回考査の入力 x_3，第 3 回考査の入力 x_4 を次で定める.

$x_1 := \{1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,$
$1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,$

$1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0.9, 0.75, 0.6, 0.6, 0.6, 0.6, 0.6, 0.5, 0.5\}$

$x_2 := \{1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0.9, 0.6, 0.6, 0.6, 0.6, 0.6\}$

$x_3 := \{1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1\}$

$x_4 := \{1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0.75, 0.5, 0.5\}$

3 層ニューラルネットワークの出力を次で定める.

$$y = w_3 tanh(w_1 x_i) + w_4 tanh(w_2 x_i).$$

また第 1 回考査について重みを次で定める.

$$w_1 = 0.512681913,\ w_3 = \frac{50.0}{56};\ w_2 = 0.153804574,\ w_4 = \frac{6.0}{56};$$

このとき $a = -0.358877,\ b = 0.785714$ である.

また第 2 回考査について重みを次で定める.

$$w_1 = 0.474230769,\ w_3 = \frac{28.0}{28};\ w_2 = 0.0,\ w_4 = \frac{0.0}{28};$$

このとき $a = -0.474231,\ b = 1$ である.

また第 3 回考査について重みを次で定める.

$$w_1 = 0.501485411,\ w_3 = \frac{43.0}{46};\ w_2 = 0.083580902,\ w_4 = \frac{3.0}{46};$$

このとき $a = -0.417905,\ b = 0.869565$ である.

このとき，パラメータ $v,\ w$ を $v_0 := 0.47423076953571436$‘, $w_0 := 1$ で固定させ，パラメータ $a,\ b$ の初期値を第 1 回から第 3 回で定める.

7.2.11　学習損失曲面上への可視化 (記号計算)

ここで理解の状況を第 1 回考査 (青)，第 2 回考査 (赤)，第 3 回考査 (緑)，第 1 回から第 3 回までを (黄) で点として表示する．　計算理解において生徒集団 (i),(iii) については図 7.6 の上側に (i),(ii) については中央に (ii),(iii) については下側に示す．

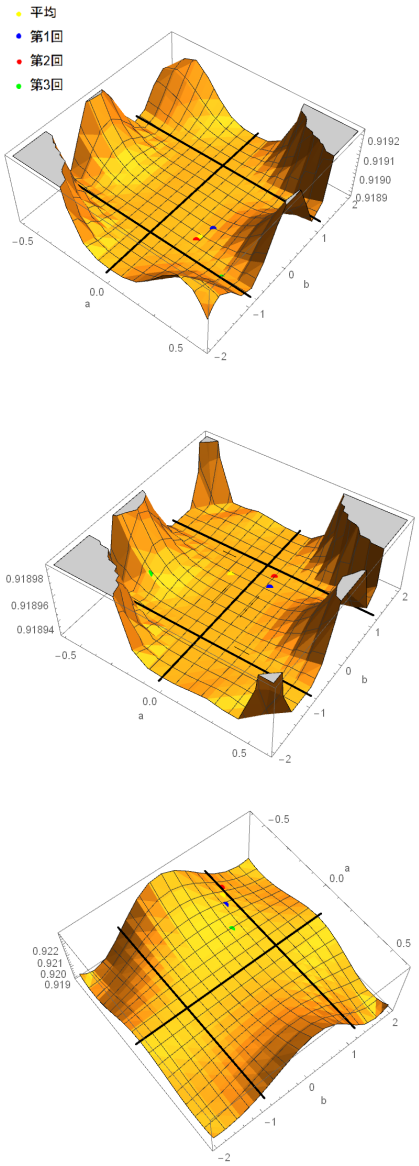


図 7.6　記号計算の学習損失曲面

第 8 章

シミュレーションの方法

試行的に作成したデータに対して各ダイナミクスの例を示し，過剰般化の程度と生徒集団の割合を変えるシミュレーションの方法について考察する．

8.1　特異領域の意味付け

定義 45，46，重みの設定 (7.1.7 参照) により，2 つの生徒集団に対して Overlap singularity 現象は「考査の平均点が等しい」, Elimination singularity 現象を「一方の生徒集団のみになる」状態とする．

ここで順列の平均点は順列を組合せとする過剰般化の程度, 組合せの平均点は組合せを順列とする過剰般化の程度を表し，「過剰般化のバランス (偏り)」は一方の集団の順列の平均点と他方の集団の組合せの平均点の差となる．

特異領域 Overlap singularity 現象の意味付けが可能である．それを次の図 8.1 で表す．

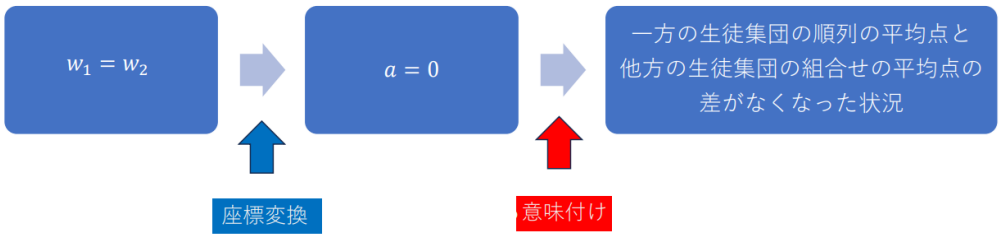


図 8.1　Overlap singularity 現象

特異領域 Elimination singularity 現象の意味付けが可能である．それを次の図 8.2 で表

す.

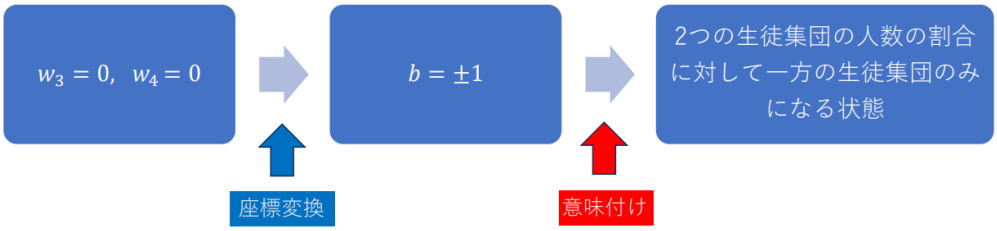


図 8.2　Elimination singularity 現象

4 つの学習段階に対して, 特異領域 Overlap singularity 現象と Elimination singularity 現象の意味付けを行う.

(1) 集団 (i) の生徒が集団 (iii) に属するために集団 (i) の組合せの平均点と集団 (iii) の順列の平均点の差がなくなった状態を Overlap singularity 現象である. また, 集団 (i) のみになる状態 ($b = 1$), または集団 (iii) のみになる状態 ($b = -1$) を Elimination singularity 現象である.

(2) 集団 (i) の生徒が集団 (ii) に属する可能性のある中で, 集団 (i) の組合せの平均点と集団 (ii) の順列の平均点の差がなくなった状況を Overlap singularity 現象である. また集団 (i) のみになる状態 ($b = 1$), または集団 (ii) のみになる状態 ($b = -1$) を Elimination singularity 現象である.

(3) 集団 (ii) の生徒が集団 (i) に属する可能性のある中で, 集団 (ii) の順列の平均点と集団 (i) の組合せの平均点の差がなくなった状態を Overlap singularity 現象である. また集団 (i) のみになる状態 ($b = 1$), または集団 (ii) のみになる状態 ($b = -1$) を Elimination singularity 現象である.

(4) 集団 (ii) の生徒が集団 (iii) に属するために集団 (i) になる可能性のある中で, 集団 (ii) の順列の平均点と集団 (iii) の組合せの平均点の差がなくなった状態を Overlap singularity 現象である. また, 集団 (iii) のみになる状態 ($b = 1$), または集団 (ii) のみになる状態 ($b = -1$) を Elimination singularity 現象である.

ここで Overlap singularity 現象, Elimination singularity 現象の対応を次の図 8.3 で表す.

特異領域	
Overlap singularity	Elimination singularity
重みパラメータ	
$\mathbf{w_1} = \mathbf{w_2}$	$w_3 = 0,\ w_4 = 0$
座標変換のパラメータ	
$a = 0$	$b = \pm 1$
数学教育における意味付け	
2つの過剰般化に差がない状態	一方の生徒集団のみになる状態

図 8.3　Overlap singularity 現象と Elimination singularity 現象

8.2　各ダイナミクスの意味付け

学習集団の理解の状況が変化したり停滞する現象を学習理論における特異領域の言葉を用いて考察する．ここで順列と組合せの問題がそれぞれ満点である生徒を完全解答と呼び，満点ではない生徒を部分点と呼ぶことにする．2 つの概念 (順列，組合せ) に対して Overlap singularity と Elimination singularity を次の現象を記述する用語として考察する．以下では生成したデータを用いてダイナミクスの挙動を説明する．

8.2.1　Overlap singularity 現象

2 つの概念 (順列，組合せ) に対して Overlap singularity 現象を次の現象を記述する用語として考察する．Overlap singularity 現象は，一方の集団の順列の平均点と他方の集団の組合せの平均点の差がなくなった状況 ($a = 0$) である (図 8.4).

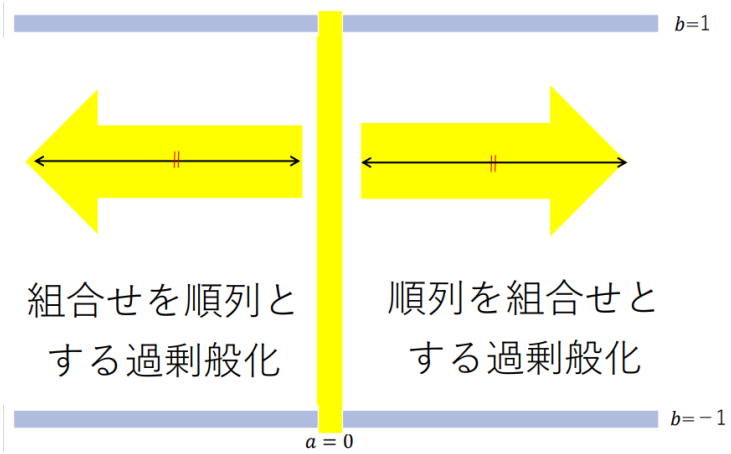


図 8.4　Overlap singularity 現象

初めに順列を学習後組合せの学習直後である場合について以下で考察する．

(1) 組合せの学習直後は組合せの学習不足のため組合せを順列とする過剰般化が起こりやすい．また組合せの学習に影響され順列を忘れてしまう生徒も存在して順列を組合せとする過剰般化が起こりやすい．よってそのバランスが取れ安定した (どちらも理解が不十分である) 状態が Overlap singularity 現象と考えられる．

例えば次のような 10 人の生徒に対して考察する．記号計算において 10 人中，順列を完全解答した生徒が 8 人 (組合せを順列と過剰般化した生徒は 8 人) とする．また組合せを完全解答した生徒は 2 人 (順列を組合せと過剰般化した生徒は 2 人) であるとする．各生徒の順列と組合せの合計点を入力として次で定める．

$$x := \{0.75, 0.75, 0.75, 0.75, 0.75, 0.75, 0.75, 0.75, 0.75, 0.75\}.$$

3 層ニューラルネットワークの重みを次で定める．

$$w_1 = 0.25,\ w_3 = \frac{8}{10};\ w_2 = 0.25,\ w_4 = \frac{2}{10}:$$

初期値を $(-0.1, 0.6)$ としてパラメータの変化と学習のダイナミクスを考え学習損失曲面を図 8.5 に示す．

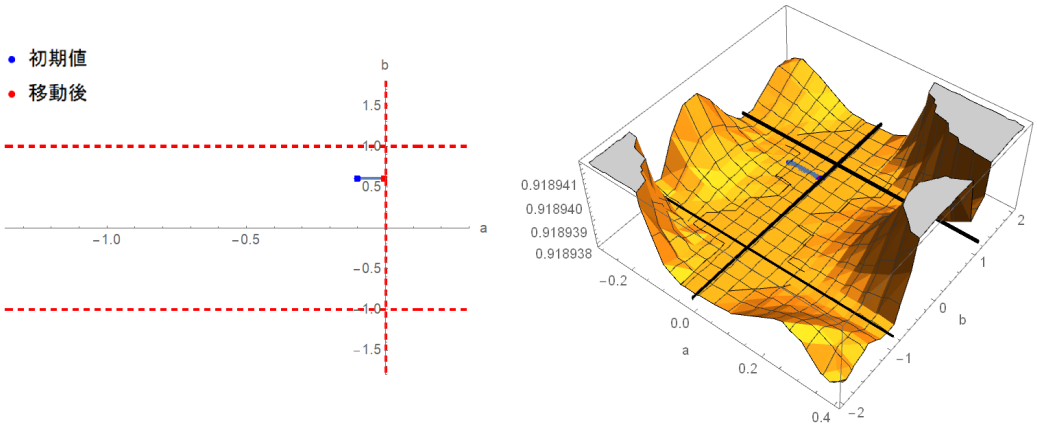


図 8.5　Overlap singularity 現象

次に組合せの学習後，時間が経過した場合について以下で考察する．

(2) 組合せを学習して時間が経過すると組合せの学習が進み順列を組合せとする過剰般化が起こりやすい．一方で順列を再学習することで組合せを順列とする過剰般化を起こす生徒も存在する．よってそのバランスが取れ安定した状態が Overlap singularity 現象と考えられる．

例えば次のような 10 人の生徒に対して考察する．記号計算において 10 人中，順列を完全解答した生徒が 2 人 (組合せを順列と過剰般化した生徒は 2 人)，また組合せを完全解答した生徒は 8 人 (順列を組合せと過剰般化した生徒は 8 人) であるとする．各生徒の順列と組合せの合計点を入力として，入力を次で定める．

$$x := \{0.75, 0.75, 0.75, 0.75, 0.75, 0.75, 0.75, 0.75, 0.75, 0.75\}.$$

3 層ニューラルネットワークの重みを次で定める．

$$w_1 = 0.25,\ w_3 = \frac{2}{10};\ w_2 = 0.25,\ w_4 = \frac{8}{10}:$$

初期値を $(0.1, -0.6)$ としてパラメータの変化と学習のダイナミクスを考え学習損失曲面を図 8.6 に示す．

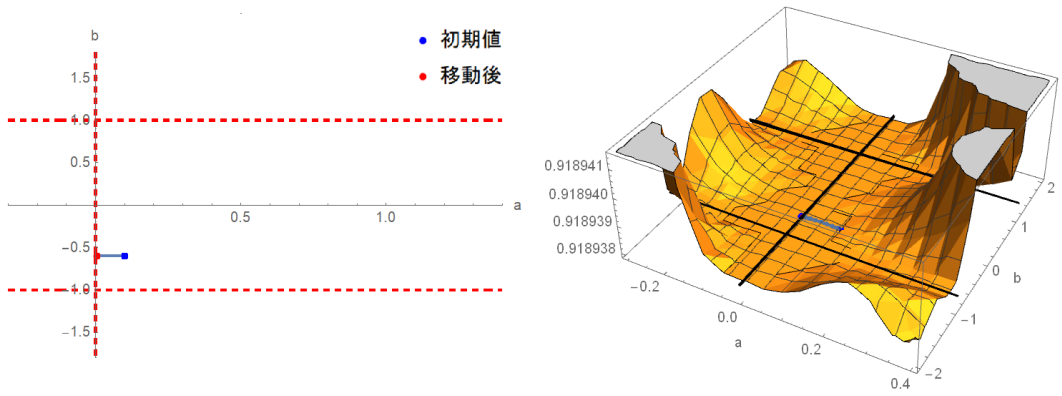


図 8.6　Overlap singularity 現象

8.2.2　Near overlap singularity 現象

2 つの概念 (順列，組合せ) に対して Near overlap singularity 現象を次の現象を記述する用語として考察する．Near overlap singularity 現象は，一方の集団の順列の平均点と他方の集団の組合せの平均点の差が大きい状態 $(a > 0)$ から差がなくなった状態 $(a = 0)$ を超えずに差が小さい状態 $(a > 0)$ へ変化する状況である（図 8.7)．

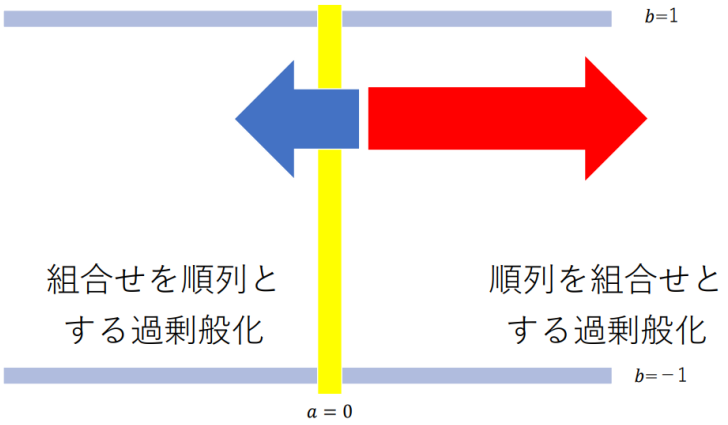


図 8.7　Near overlap singularity 現象

例えば次のような 10 人の生徒に対して考察する．記号計算において 10 人中，順列を完全解答した生徒が 0 人 (組合せを順列と過剰般化した生徒は 0 人)，また組合せを完全解答した生徒は 10 人 (順列を組合せと過剰般化した生徒は 10 人) であるとする．各生徒の順列と組合せの合計点を入力として，次で定める．

$$x := \{0.75, 0.75, 0.75, 0.75, 0.75, 0.75, 0.75, 0.75, 0.75, 0.75\}.$$

3 層ニューラルネットワークの重みを次で定める．

$$w_1 = 0.0,\ w_3 = \frac{0}{10};\ w_2 = 0.25,\ w_4 = \frac{10}{10}:$$

初期値を $(0.25, -0.8)$ としてパラメータの変化と学習のダイナミクスを考え学習損失曲面を図 8.8 に示す．

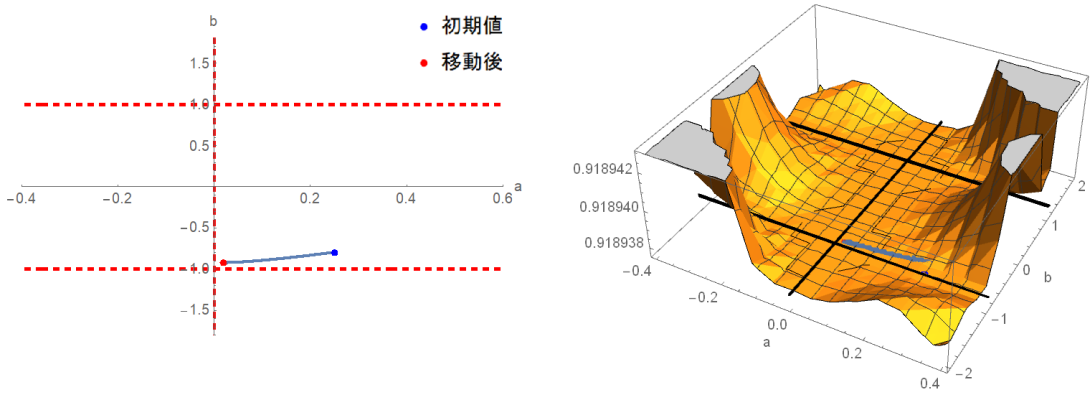


図 8.8　Near overlap singularity 現象

$a = 0$ に近づき Near overlap singularity の現象が起こる．

8.2.3 Cross overlap singularity 現象

2 つの概念 (順列，組合せ) に対して Cross overlap singularity 現象を次の現象を記述する用語として考察する．Cross overlap singularity 現象は，一方の集団の順列の平均点と他方の集団の組合せの平均点の差が大きい状態 $(a > 0)$ から差がなくなった状態 $(a = 0)$ を超えて他方の集団の組合せの平均点と一方の集団の順列の平均点の差が大きい状態 $(a < 0)$ へ変化する状況である（図 8.9)．

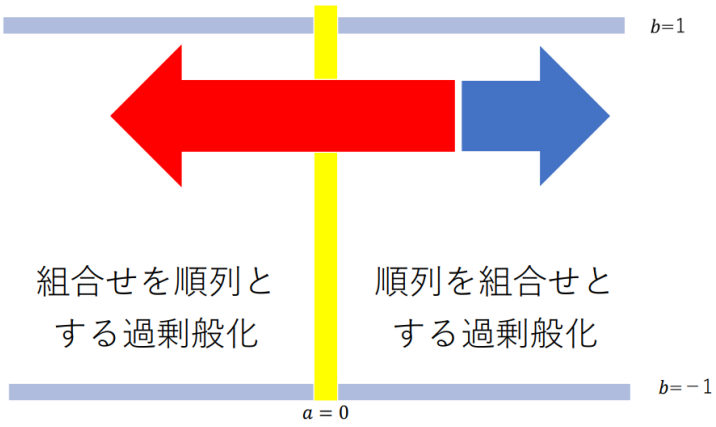


図 8.9　Cross overlap singularity 現象

順列を学習後，組合せの学習直後である場合について以下で考察する．

意味理解において組合せを学習した直後は順列を完全に理解している生徒が多く，組合せを順列とする過剰般化が大きい．組合せを順列とする過剰般化が大きい状態から，逆に順列を組合せとする過剰般化が大きくなる現象が起こると，Cross overlap singularity 現象として考えられる．

例えば次のような 8 人の生徒に対して考察する．意味理解において 8 人中，順列を完全解答した生徒が 1 人 (組合せを順列と過剰般化した生徒は 1 人)，組合せを完全解答した生徒は 7 人 (順列を組合せと過剰般化した生徒は 7 人) であるとする．各生徒の順列と組合せの合計点を入力として，次で定める．

$$x := \{0.75, 0.75, 0.75, 0.75, 0.75, 0.8, 0.8, 0.8\}.$$

3 層ニューラルネットワークの重みを次で定める．

$$w_1 = 0.25,\ w_3 = \frac{1}{8};\ w_2 = 0.27,\ w_4 = \frac{7}{8}:$$

初期値を $(0.4, 0.6)$ としてパラメータの変化と学習のダイナミクスを考え学習損失曲面を図 8.10 に示す.

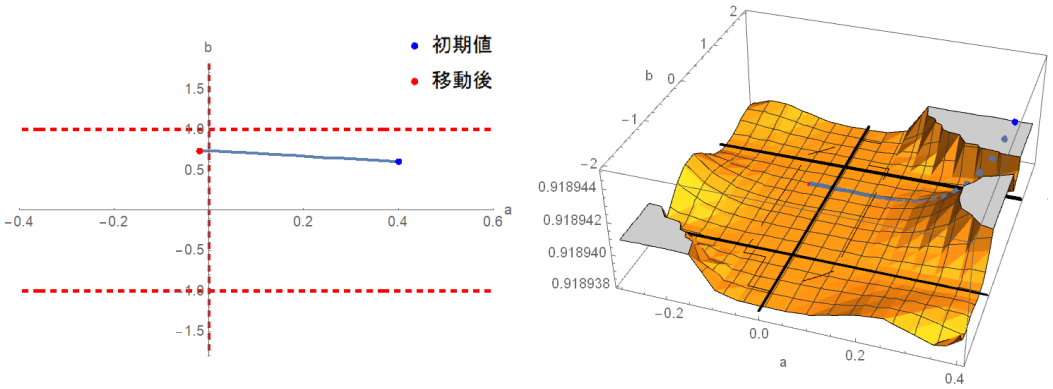


図 8.10　Cross overlap singularity 現象

$a = 0$ を通過して Cross overlap singularity 現象が起こる.

8.2.4　Elimimation singularity 現象

2 つの概念 (順列，組合せ) に対して Elimination singularity は次の現象を記述する用語として考察する．Elimination singularity 現象は，2 つの生徒集団に対して順列が完全解答である生徒集団のみになる状態 $(b = 1)$ または組合せが完全解答である生徒集団のみになる状態 $(b = -1)$ 状況である（図 8.11).

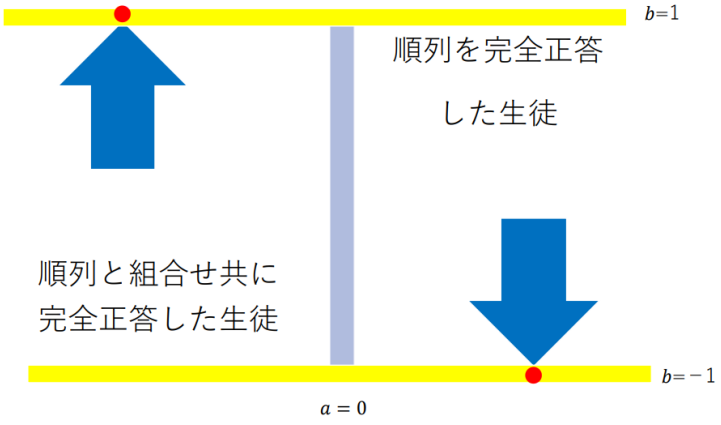


図 8.11　Elimination singularity 現象

初めに順列を学習後，組合せの学習直後である場合について以下で考察する．

(1) 組合せの学習直後は順列を完全に理解した生徒のみで，組合せを順列とする過剰般化を起こす生徒が存在する．組合せを完全に理解した生徒がおらず順列を組合せとする過剰般化が起こらない状態であり，順列の理解が進み (正の転移)，組合せの理解に影響を与えている (負の転移)．よってその偏りがある状態が Elimination singularity 現象と考えられる．

例えば次のような 10 人の生徒に対して考察する．記号計算において 10 人中，順列を完全解答した生徒が 10 人 (組合せを順列と過剰般化した生徒は 10 人) とする． また組合せを完全解答した生徒は 0 人 (順列を組合せと過剰般化した生徒は 0 人) であるとする．各生徒の順列と組合せの合計点を入力として，入力を次で定める．

$$x := \{0.75, 0.75, 0.75, 0.75, 0.75, 0.75, 0.75, 0.75, 0.75, 0.75\}.$$

3 層ニューラルネットワークの重みを次で定める．

$$w_1 = 0.25,\ w_3 = \frac{10}{10};\ w_2 = 0.0,\ w_4 = \frac{0}{10}\ :$$

初期値を $(-0.3, 0.8)$ としてパラメータの変化と学習のダイナミクスを考え学習損失曲面を図 8.12 に示す．

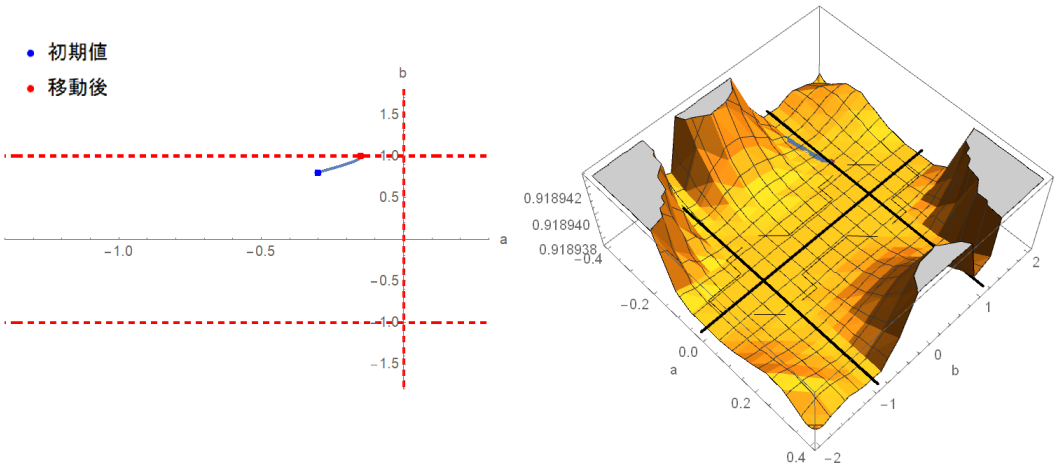


図 8.12　Elimination singularity 現象

$b = 1$ に近づき Elimination singularity 現象が起こる．順列を完全解答した生徒のみになり，組合せを完全解答した生徒がいなくなる．

次に組合せの学習後，時間が経過した場合について以下で考察する．

(2) 組合せを学習して時間が経過すると組合せを完全に理解した生徒のみで，順列を組合せとする過剰般化を起こす生徒が存在する．順列を完全に理解した生徒がおらず組合せを順列とする過剰般化が起きない状態であり，組合せの理解が進み (正の転移)，順列の理解に影響を与えている (負の転移)．よってその偏りがある状態が Elimination singularity 現象と考えられる．

例えば次のような 10 人の生徒に対して考察する．記号計算において 10 人中，順列を完全解答した生徒が 0 人 (組合せを順列と過剰般化した生徒は 0 人)，また組合せを完全解答した生徒は 10 人 (順列を組合せと過剰般化した生徒は 10 人) であるとする．各生徒の順列と組合せの合計点を入力として，次で定める．

$$x := \{0.75, 0.75, 0.75, 0.75, 0.75, 0.75, 0.75, 0.75, 0.75, 0.75\}.$$

3 層ニューラルネットワークの重みを次で定める．

$$w_1 = 0.0,\ w_3 = \frac{0}{10};\ w_2 = 0.25,\ w_4 = \frac{10}{10}:$$

初期値を $(0.3, -0.8)$ としてパラメータの変化と学習のダイナミクスを考え学習損失曲面を図 8.13 に示す．

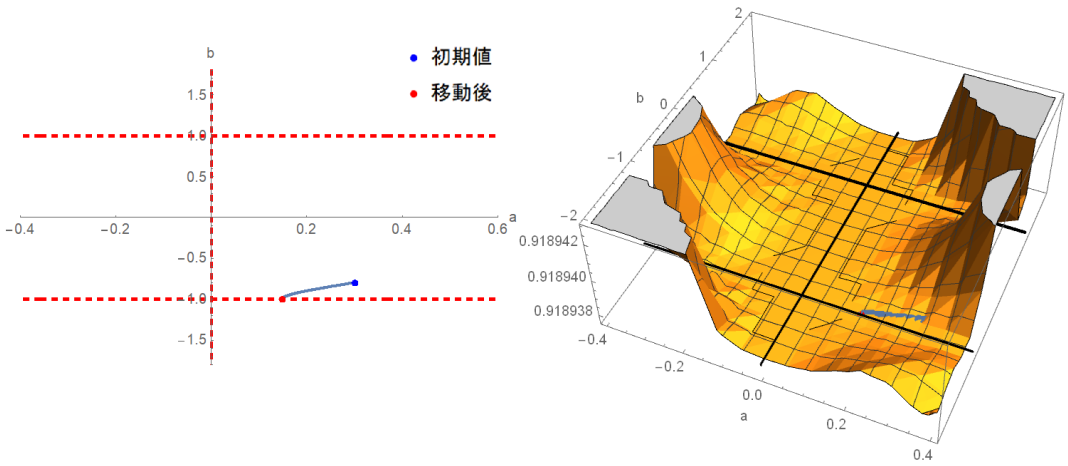


図 8.13　Elimination singularity 現象

$b = -1$ に近づき Elimination singularity 現象が起こる．組合せを完全解答した生徒のみになり，順列を完全解答した生徒がいなくなる．

8.2.5 Near elimination singularity 現象

2 つの概念 (順列，組合せ) に対して Near elimination singularity を次の現象を記述する用語として考察する．Near elimination singularity 現象は，2 つの生徒集団に対して，順列が完全解答である生徒集団のみになる状態 ($b = 1$) または組合せが完全解答である生徒集団のみになる状態 ($b = -1$) に近づく状況である（図 8.14)．

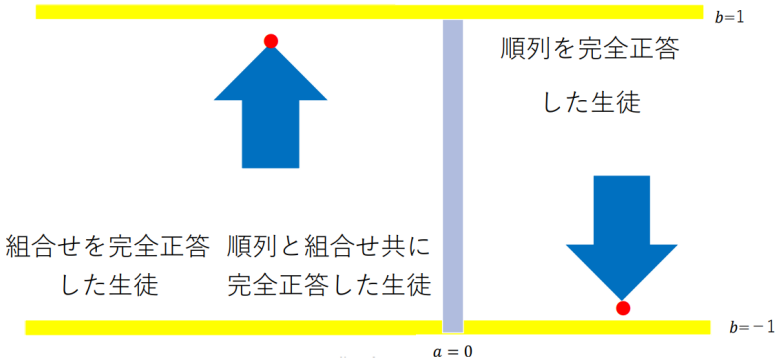


図 8.14　Near elimination singularity 現象

例えば次のような 10 人の生徒に対して考察する．記号計算において 10 人中，順列を完全解答した生徒が 3 人 (組合せを順列と過剰般化した生徒は 3 人) とする． また組合せを完全解答した生徒は 7 人 (順列を組合せと過剰般化した生徒は 7 人) であるとする．各生徒の順列と組合せの合計点を入力として，次で定める．

$$x := \{0.65, 0.65, 0.65, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8, 0.8\}.$$

3 層ニューラルネットワークの重みを次で定める．

$$w_1 = 0.15,\ w_3 = \frac{3}{10};\ w_2 = 0.3,\ w_4 = \frac{7}{10}\ :$$

初期値を $(-0.5, -0.8)$ としてパラメータの変化と学習のダイナミクスを考え学習損失曲面を図 8.15 に示す．

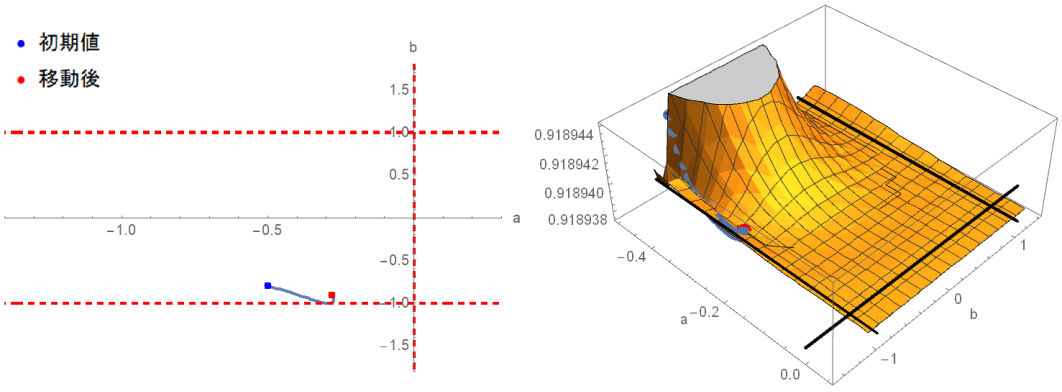


図 8.15　Near elimination singularity 現象

$b = -1$ に近づき Near elimination singularity 現象が起こる.

8.2.6　Fast convergence 現象

2 つの概念 (順列，組合せ) に対して Fast convergence を次の現象を記述する用語として考察する．Fast convergence 現象は，一方の集団の順列の平均点と他方の集団の組合せの平均点の差が小さい状態 $(a < 0)$ から差がなくなった状態 $(a = 0)$ を超えずに差が大きい状態 $(a < 0)$ へ変化する状況である（図 8.16).

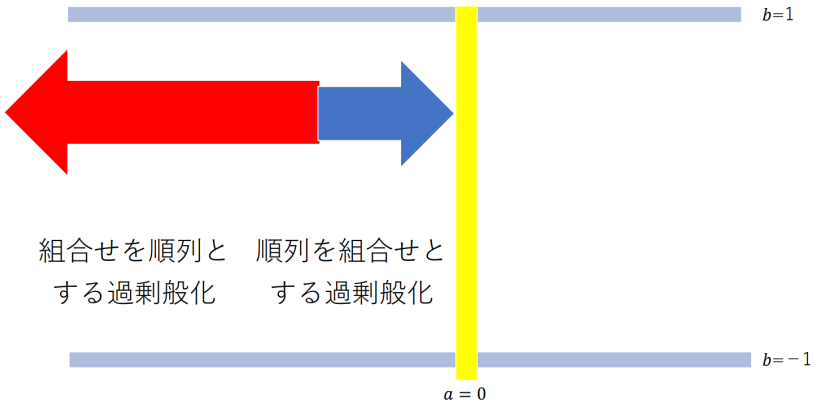


図 8.16　Fast convergence 現象

例えば次のような 10 人の生徒に対して考察する.

記号計算において 10 人中，順列を完全解答した生徒が 3 人 (組合せを順列と過剰般化

した生徒は 0 人)，また組合せを完全解答した生徒は 7 人 (順列を組合せと過剰般化した生徒は 7 人) であるとする．各生徒の順列と組合せの合計点を入力として，次で定める．

$$x := \{1, 1, 1, 0.65, 0.65, 0.65, 0.65, 0.65, 0.65, 0.65\}.$$

3 層ニューラルネットワークの重みを次で定める．

$$w_1 = 0.5;\ w_3 = \frac{3}{10};\ w_2 = 0.15;\ w_4 = \frac{7}{10}:$$

初期値を $(-0.1, -0.4)$ としてパラメータの変化と学習のダイナミクスを考え学習損失曲面を図 8.17 に示す．

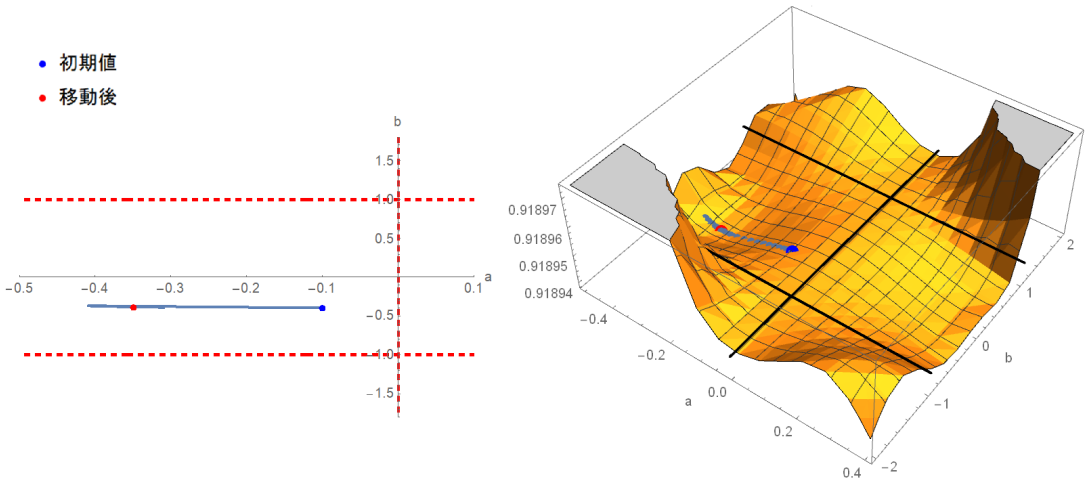


図 8.17　Fast convergence 現象

$a = 0$ から遠ざかり Fast convergence 現象が起こる．

8.3　シミュレーション

データをもとに得た結果を用いて，a については -0.6 から 0.6 まで 0.05 ずつ変化させて，b については -1.1 から 1.1 まで 0.05 ずつ変化させてシミレーションを行う．ここで初期値について第 1 回考査 (青)，第 2 回考査 (赤)，第 3 回考査 (緑) を◎，真の分布を $\times$ で表示する．第 1 回から第 2 回考査へのダイナミクスについては初期値 (シミュレーション) を青，その移動後を赤で表す．第 2 回から第 3 回考査へのダイナミクスについては初期値 (シミュレーション) を赤，移動後を緑で表す．

過剰般化の程度を変えることで生徒集団の人数の割合の変化から理解が進んでいるか判断できる．生徒の割合が変化すると理解が進む (進んでいない) ので理解の変容はありと判断でき，変わらなければ理解が停滞するので理解の変容はないと判断できる（図 8.18）．

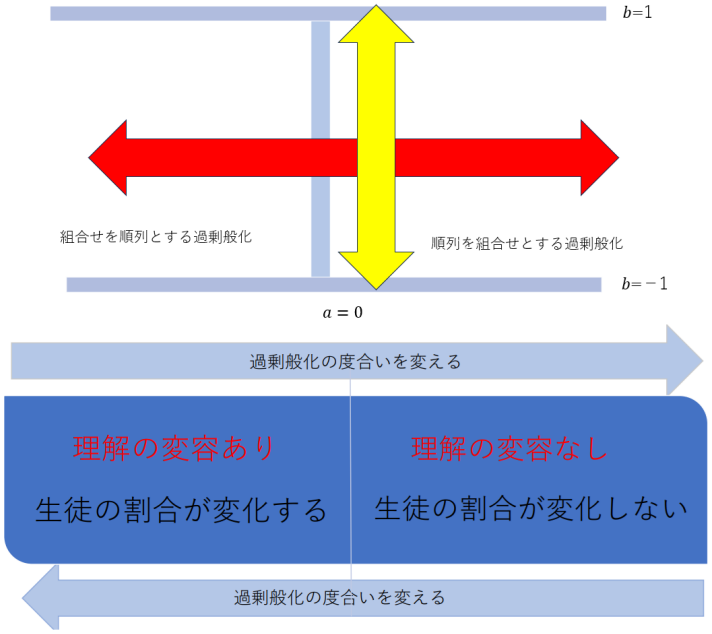


図 8.18　過剰般化の程度を変える

2つの生徒集団の人数の割合を変えることで，過剰般化が抑えられる (正の転移) か広がる (負の転移) か判断できる．生徒集団の人数の割合を変えて過剰般化の程度を調べる．過剰般化が抑えれれば理解が進むので正の転移，過剰般化が広がれば (または逆の過剰般化が大きくなれば) 理解が進まないので負の転移であると判断できる（図 8.19)．

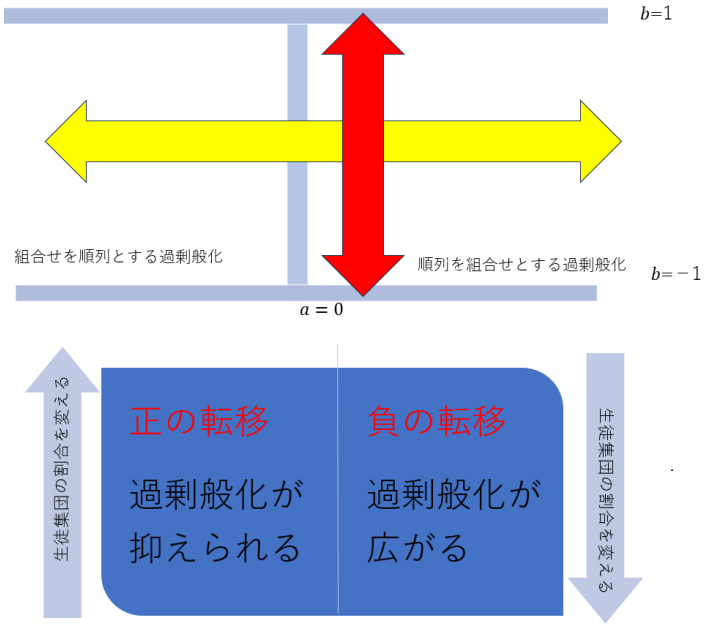


図 8.19　生徒集団の割合を変える

第 9 章

テストデータでの分析

意味理解と記号計算からなるテストデータについてそれぞれ 4 つの学習段階に分けて a 変化と b 変化のシミュレーションを行う.

9.1　意味理解における分析

9.1.1　相互関係の問題を正答した生徒 (意味理解 : 第 1 回から第 2 回)

組合せの学習直後に「1 人 1 人の役職を選ぶ」問題を順列で正解できた生徒を対象とするため第 1 回から第 2 回考査の変化を分析する. このとき順列は完全正答して組合せは部分点の生徒と順列と組合せ共に完全正答の生徒を比べて組合せの理解について考察する.

a 変化のシミュレーション

a を変化させたシミュレーションのダイナミクスと訓練損失の変化と損失曲面上のダイナミクスを図 9.1 に示す.

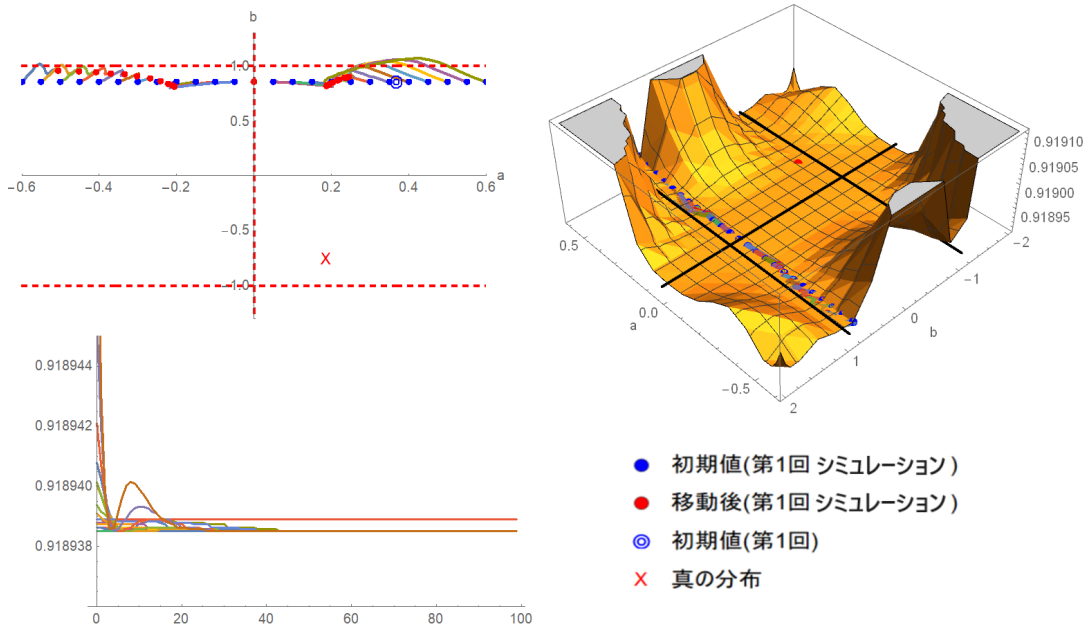


図 9.1　意味:a 変化: 正答

順列を完全正答して組合せが部分点である生徒の割合の方が多く偏りがある状態 ($b = 0.85$) から学習を始める.

初めに組合せを順列とする過剰般化が大きくなる (a が減少する) と，臨界直線 $b = 1$ の影響を受けて Elimination singularity 現象が起こる．順列を完全正答して組合せは部分点の生徒のみになり，順列と組合せ共に完全正答した生徒がいなくなる．組合せを順列とする過剰般化が $a < -0.3$ より小さくなると，組合せの学習が進まずに順列の理解の変容が起こる.

一方で順列を組合せとする過剰般化が大きくなる (a が増加する) と，臨界直線 $b = 1$ の影響を受けて，近づくが元に戻り $a = 0.3$ で停滞して，Near elimination singularity 現象が起きる．順列を完全正答して組合せは部分点の生徒と，順列と組合せともに完全正答した生徒の割合は初期状態と変わらず，順列の理解の変容が起こらない.

b 変化のシミュレーション

b を変化させたシミュレーションのダイナミクスと訓練損失の変化と損失曲面上のダイナミクスを図 9.2 に示す.

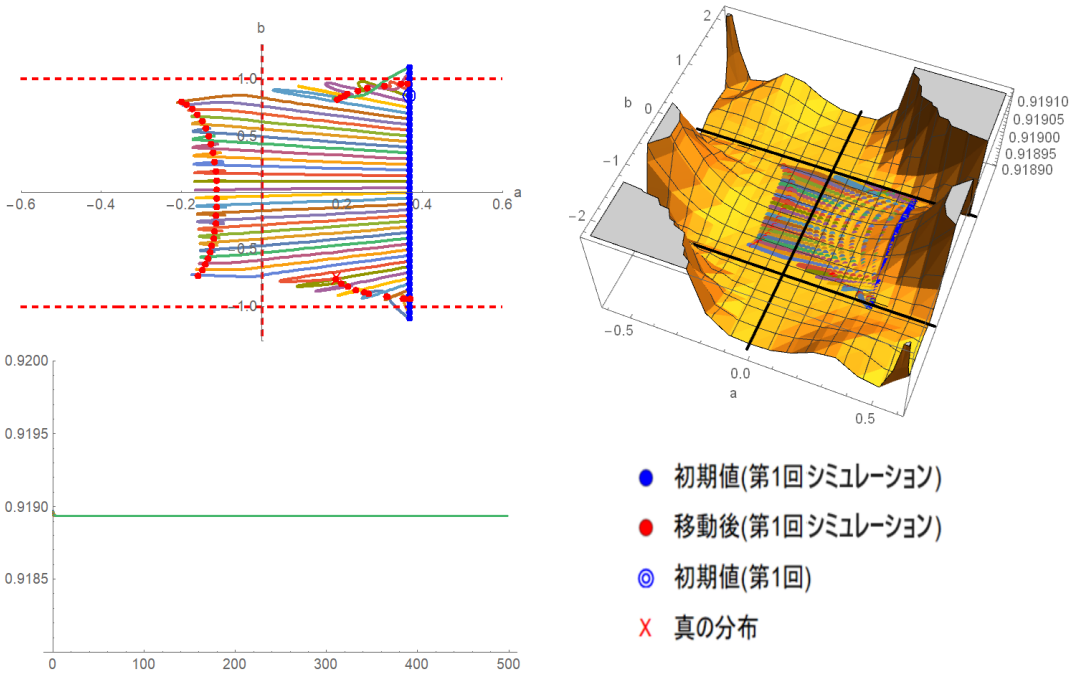


図 9.2　意味:b 変化: 正答

順列を組合せとする過剰般化が大きい状態 ($a = 0.36$) から学習を始める.

初めに順列は完全正答して組合せが部分点の生徒の割合を増やす (b が増加する) と，臨界直線 $b = 1$ 近づき元に戻る Near elimination singularity 現象が起こる．順列を組合せとする過剰般化が大きい状態が抑えられ正の転移が起こる.

また順列と組合せをともに完全正答した生徒の割合を増やす (b が減少する) と，臨界直線 $a = 0$ を超えて Cross overlap singularity 現象が起こる．初めは順列を組合せとする過剰般化が大きい状態から，学習の過程で $v = 0.35$ より w_1, w_2 が 0.35 で過剰般化がなくなる状態 (Overlap singularity 現象) になる．さらに組合せを順列とする過剰般化が進み，組合せの学習が進まずに負の転移が起きる.

9.1.2　相互関係の問題を準正答した生徒 (意味理解 : 第 1 回から第 2 回)

組合せの学習直後に，「1 人 1 人の役職を選ぶ」問題を組合せを用いて解答し，準正答した生徒を対象とするため第 1 回から第 2 回考査の変化を分析する．

このとき順列は完全正答して組合せは部分点の生徒と，組合せは完全正答して順列は部分点の生徒を比べて 2 つの概念を理解していく過程について考察する．

a 変化のシミュレーション

a を変化させたシミュレーションのダイナミクスと訓練損失の変化と損失曲面上のダイナミクスを図 9.3 に示す．

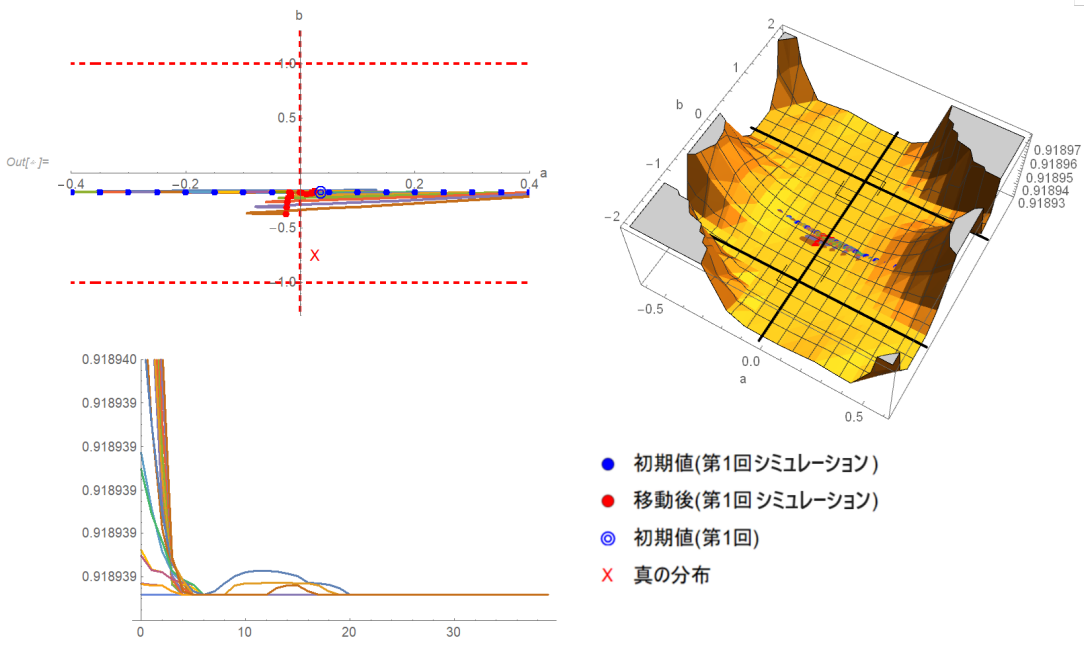


図 9.3　意味:a 変化: 準正答

順列を完全正答して組合せが部分点である生徒より，組合せは完全正答して順列は部分点である生徒の割合が少し多い状態 ($b = -0.17$) から学習を始める．

組合せを順列とする過剰般化が大きくなる (a が増加する) と，臨界直線 $a = 0$ の影響を受けて $a = 0.025$ で停滞して Cross overlap singularity 現象が起こる．順列を完全正答して組合せは部分点の生徒と，組合せを完全正答して順列は部分点である生徒の割合は初期状態と変わらないので組合せの学習が進まずに，順列の理解の変容が起こらない．

また，順列を組合せとする過剰般化が大きくなる (a が減少する) と，臨界直線 $a = 0$ の影響を受けて $a = -0.019$ で停滞して Cross overlap singularity 現象が起こる．組合せは完全正答して順列が部分点である生徒の割合が少し増えるため組合せの学習が少し進み，順列の理解の変容が起こる．

b 変化のシミュレーション

b を変化させたシミュレーションのダイナミクスと訓練損失の変化と損失曲面上のダイナミクスを図 9.4 に示す．

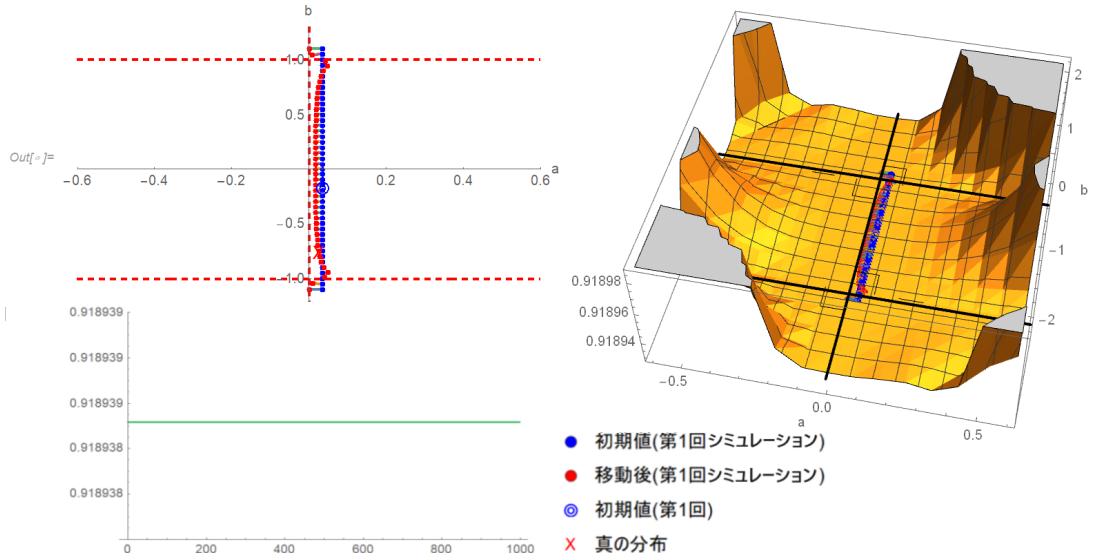


図 9.4　意味:b 変化: 準正答

順列を組合せとする過剰般化が大きい状態 ($a = 0.035$) から学習を始める．

順列は完全正答して組合せが部分点の生徒の割合を増やす (b が増加する) と，臨界直線 $a = 0$ の影響を受けて Overlap singularity 現象が起こる．順列を組合せとする過剰般化が大きい状態から，学習の過程で $v = 0.26$ より w_1，w_2 が 0.26 で組合せを順列とする過剰般化との差がない状態になるため正の転移が起こる．

また，組合せは完全正答して順列が部分点の生徒の割合を増やす (b が減少する) 場合も同様であり，正の転移が起こる．

9.1.3　相互関係の問題を準正答した生徒 (意味理解 : 第 2 回から第 3 回)

組合せの学習後時間が経過した後に「1 人 1 人の役職を選ぶ」問題を組合せを用いてしまい準正解である生徒を対象とするため第 2 回から第 3 回考査の変化を分析する．このとき順列は完全正答して組合せは部分点の生徒と組合せは完全正答して順列は部分点の生徒を比べて順列又は組合せが部分点である生徒について考察する．

a 変化のシミュレーション

a を変化させたシミュレーションのダイナミクスと訓練損失の変化と損失曲面上のダイナミクスを図 9.5 に示す．

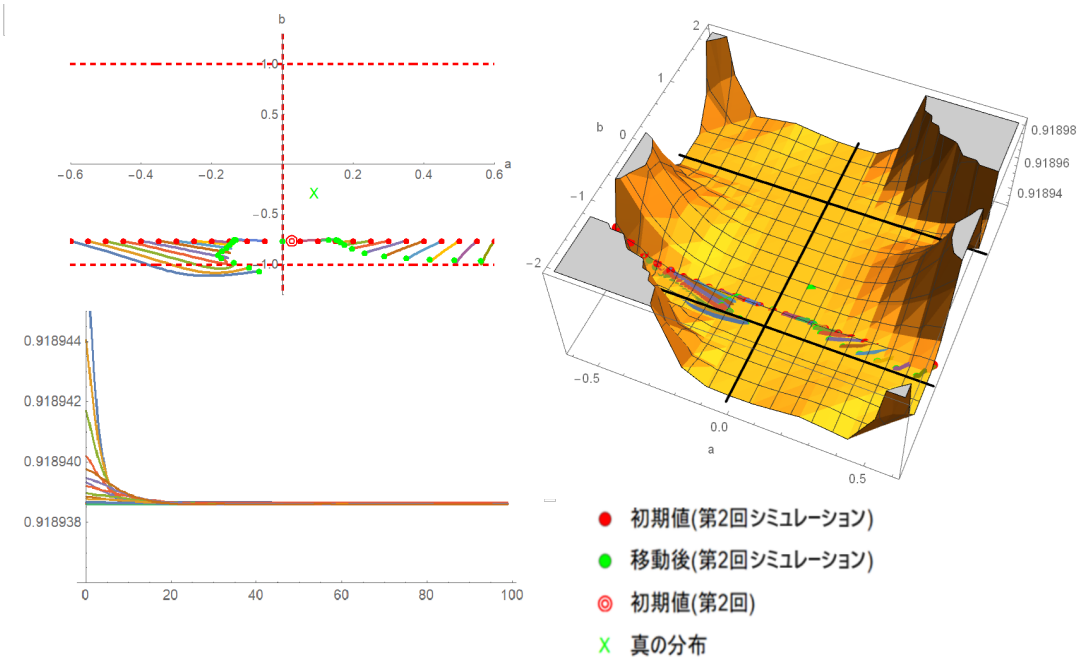


図 9.5　意味:a 変化: 準正答

組合せは完全正答して順列は部分点の生徒の割合が多い状態 ($b = -0.76$) から学習を始める．

初めに順列を組合せとする過剰般化が大きくなる (a が増加する) と，臨界直線 $b = -1$ の影響を受けて Elimination singularity 現象が起こる．組合せは完全正答して順列が部

分点である生徒のみになり，順列は完全正答して組合せが部分点である生徒がいなくなる．順列を組合せとする過剰般化が $a > 0.3$ より大きくなると，順列の再学習が進まずに組合せの理解の変容が起こる．

また組合せを順列とする過剰般化が大きくなる (a が減少する) と，臨界直線 $b = -1$ の影響を受けて，近づくが元に戻り，$a = -0.15$ で停滞して，Near elimination singularity 現象が起きる．順列を完全正答して組合せは部分点の生徒と，組合せは完全正答して順列が部分点である生徒の割合は初期状態と変わらず，組合せの理解の変容が起こらない．

b 変化のシミュレーション

b を変化させたシミュレーションのダイナミクスと訓練損失の変化と損失曲面上のダイナミクスを図 9.6 に示す．

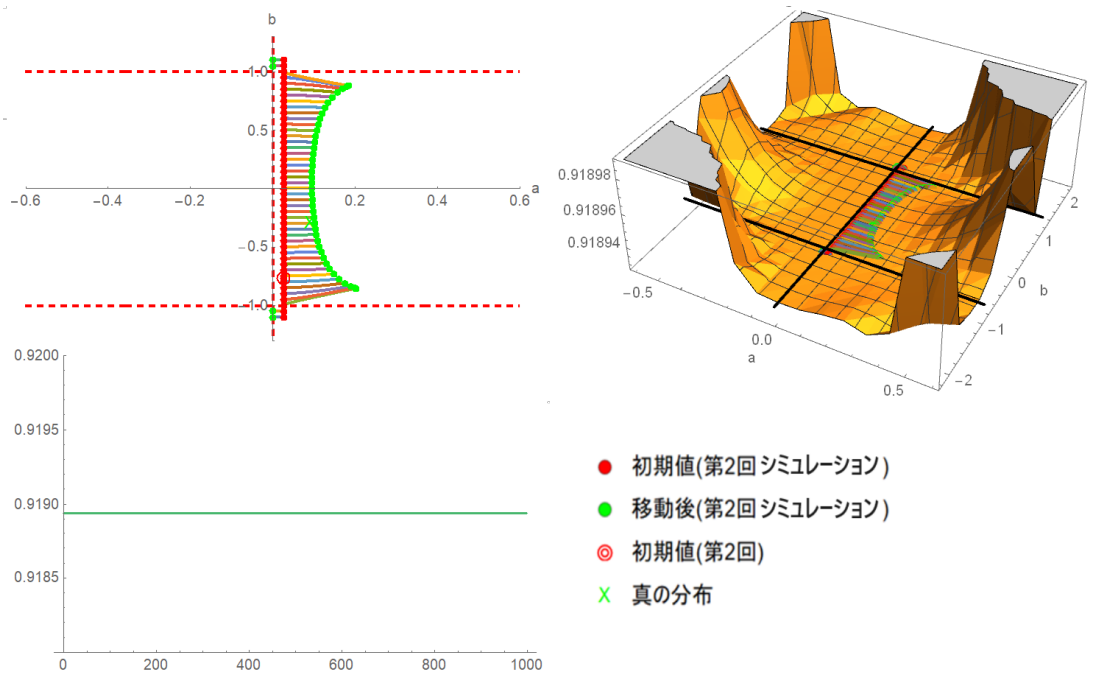


図 9.6　意味:b 変化: 準正答

順列を組合せとする過剰般化が少し大きい状態 ($a = 0.026$) から学習を始める．初めに順列は完全正答して組合せが部分点である生徒の割合を増やす (b が増加する) と，臨界直線 $a = 0$ の影響を受けずに Fast convergence 現象が起こる．順列を組合せとする過剰般

化が少し小さくなり正の転移が起こる．

また組合せは完全正答して順列が部分点である生徒の割合を増やす (b が減少する) と，臨界直線 $a = 0$ の影響を受けずに Fast convergence 現象が起こる．順列を組合せとする過剰般化がさらに大きくなり負の転移が起こる．学習の過程で順列を組合せとする過剰般化が大きい状態からさらに順列を組合せとする過剰般化が進み，負の転移が起こる．過剰般化が起こった状態であると考えることができる．順列の再学習をすることで順列を組合せとする過剰般化は $a = 0.18$ から $a = 0.094$ に抑えることができる．

9.1.4 相互関係の問題を正答した生徒 (意味理解 : 第 2 回から第 3 回)

組合せの学習後に時間が経過した後に，「1 人 1 人の役職を選ぶ」問題において，順列の再学習後に相互関係を考慮して正解できる生徒を対象とするために第 2 回から第 3 回考査の変化を分析する．このとき，組合せは完全正答して順列は部分点の生徒と，順列と組合せともに完全正答した生徒を比べて，2 つの概念を理解していく過程について考察する．

a 変化のシミュレーション

初めに a を変化させたシミュレーションのダイナミクスと訓練損失の変化と損失曲面上のダイナミクスを図 9.7 に示す.

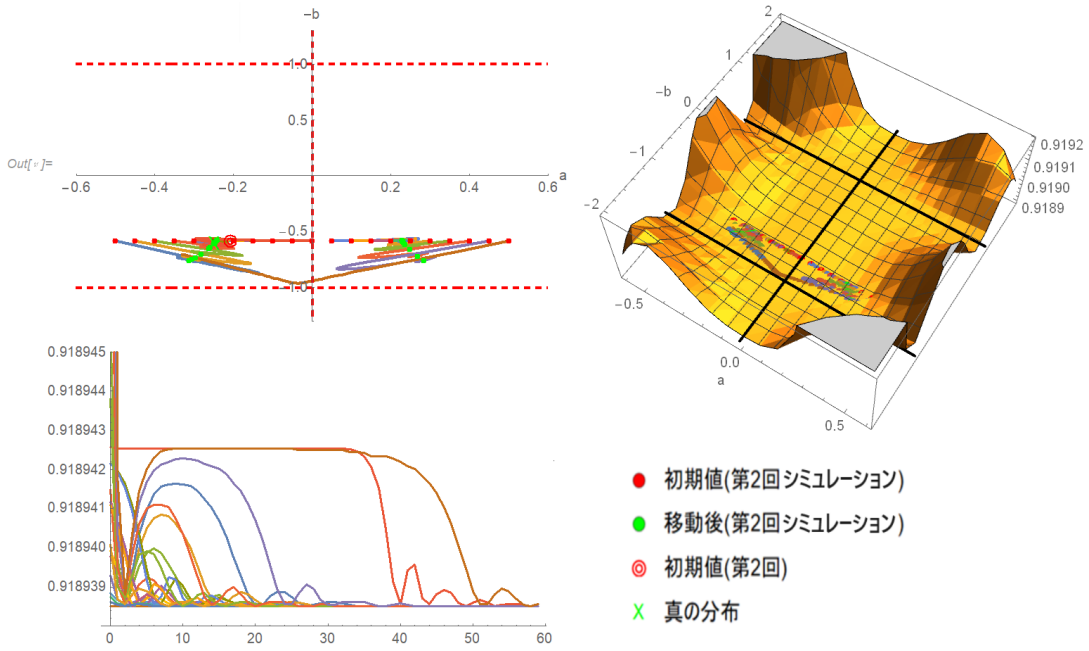


図 9.7　意味:a 変化: 正答

順列と組合せともに完全正答した生徒の割合が多い状態 ($b = 0.58$) から学習を始める.

組合せを順列とした過剰般化が大きくなる (a が増加する) と，臨界直線 $a = 0$ の影響を受けて $a = 0.25$ で停滞し，Near overlap singularity 現象が起こる．順列と組合せともに完全正答した生徒の割合が $b = 0.58$ から $b = 0.75$ まで少し増えるため，順列の再学習が少し進み，組合せの理解の変容が起こる.

また，順列を組合せとする過剰般化が大きくなる (a が減少する) と，$a = -0.25$ で停滞して同様の状態になり，順列の再学習が少し進み，組合せの理解の変容が起こる.

b 変化のシミュレーション

次に b を変化させたシミュレーションのダイナミクスと，訓練損失の変化と損失曲面上のダイナミクスを図 9.8 に示す．

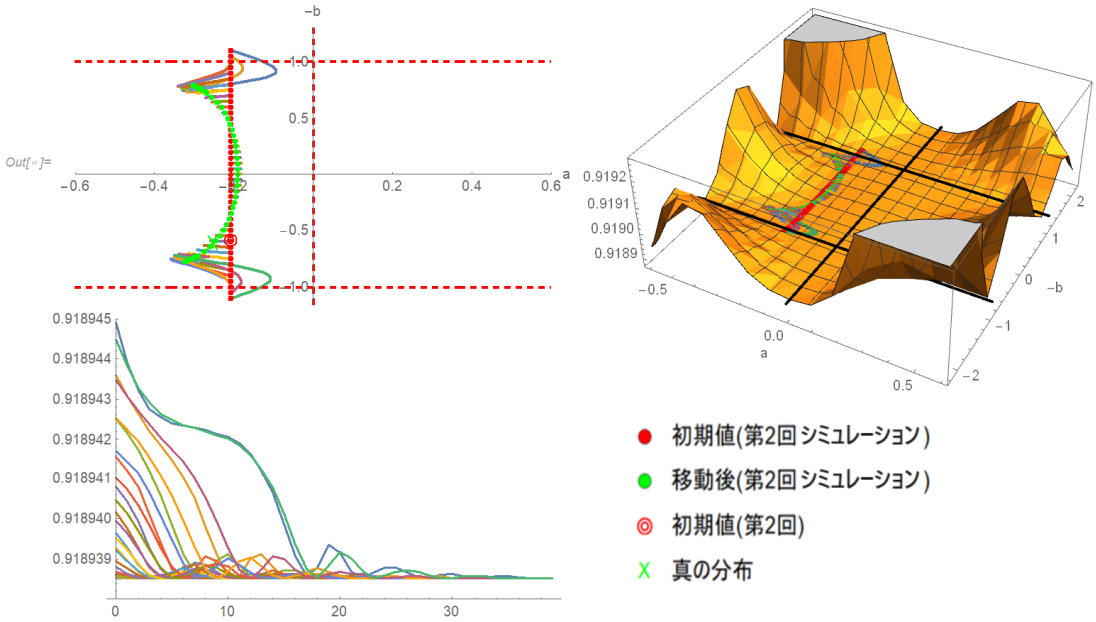


図 9.8　意味:b 変化: 正答

組合せを順列とする過剰般化が大きく偏りがある状態 ($a = -0.208$) から学習を始める．

順列と組合せをともに完全正答する生徒の割合を増やす (b が増加する) と，$b < -0.55$ ，$b > 0.4$ のとき，a は小さくなり臨界直線 $a = 0$ の影響を受けずに収束する (Fast convergence 現象)．$b = 0.4$ から $b = 0.76$ まではさらに組合せを順列とする過剰般化が大きくなるため負の転移が起こる．

また，組合せは完全正答で順列が部分点である生徒の割合を増やす (b が減少する) と，$-0.55 \leq b \leq 0.4$ のとき，a は大きくなり Near overlap singularity 現象が起こる．組合せを順列とする過剰般化が小さくなるため正の転移が起こる．

9.2　記号計算における分析

9.2.1　相互関係の問題を正答した生徒 (記号計算 : 第 1 回から第 2 回)

組合せの学習直後に「1 人 1 人の役職を選ぶ」問題を順列で正解できた生徒を対象として第 1 回から第 2 回考査の変化を分析する．このとき順列は完全正答して組合せは部分点の生徒と順列と組合せ共に完全正答の生徒を比べて組合せの理解について考察する．

a 変化のシミュレーション

a を変化させたシミュレーションのダイナミクスと訓練損失の変化と損失曲面上のダイナミクスを図 9.9 に示す．

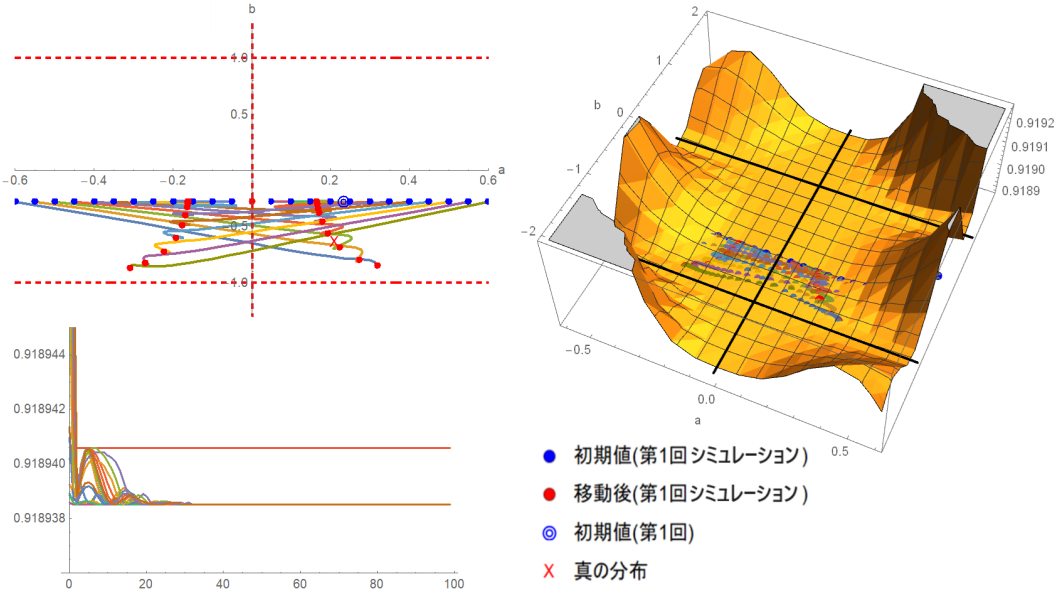


図 9.9　記号:a 変化: 正答

順列と組合せともに完全正答した生徒の割合が少し多い状態 ($b = -0.28$) から学習を始める．

組合せを順列とした過剰般化が大きくなる (a が減少する) と，臨界直線 $a = 0$ の影響を受けて $a = -0.16$ で停滞し，Near overlap singularity 現象が起こる．さらに $v = 0.45$ より w_1, w_2 が 0.45 で過剰般化がなくなる状態 (Overlap singularity 現象) になり，最後は，$a = 0.16$ で停滞し，Cross overlap singularity 現象が起こる．順列と組合せともに完全正答した生徒の割合が $b = -0.28$ から $b = -0.83$ 減るため，組合せの学習が進み，順列の理解の変容が起こる．

また，順列を組合せとする過剰般化が大きくなる (a が増加する) と，同様の状態になり，組合せの学習が進み，順列の理解の変容が起こる．

b 変化のシミュレーション

b を変化させたシミュレーションのダイナミクスと訓練損失の変化と損失曲面上のダイナミクスを図 9.10 に示す．

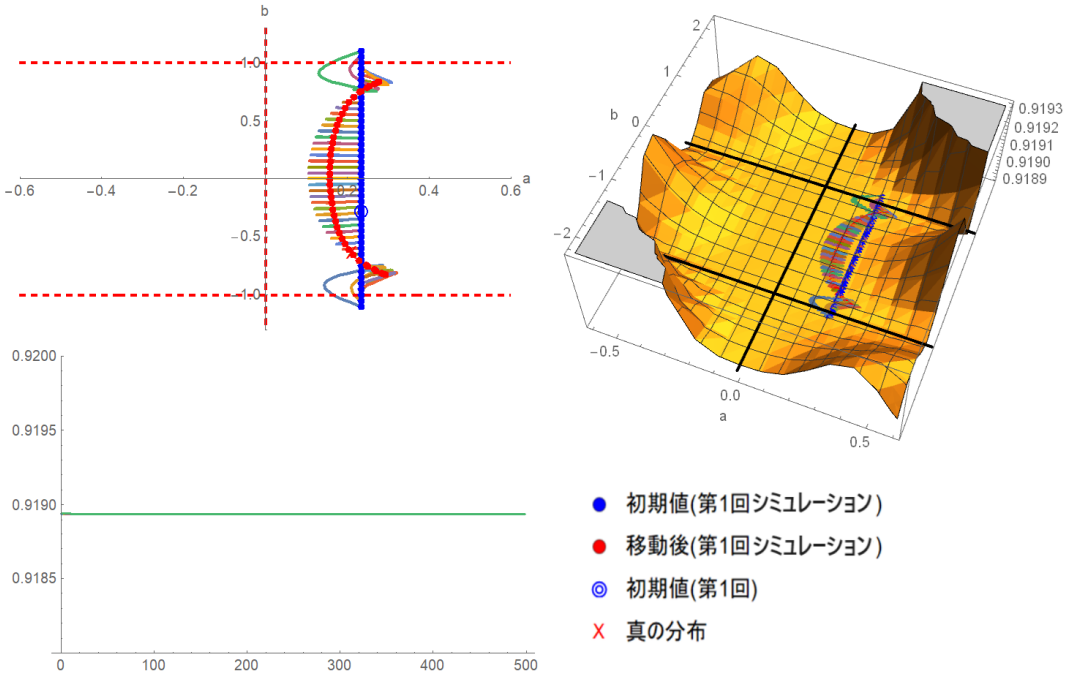


図 9.10　記号:b 変化: 正答

順列を組合せとする過剰般化が大きく偏りがある状態 ($a = 0.233$) から学習を始める．

順列と組合せをともに完全正答する生徒の割合を増やす (b が減少する) と，$b < -0.74$, $b > 0.74$ のとき，a は大きくなり臨界直線 $a = 0$ の影響を受けずに収束 (Fast convergence 現象) する．$b = -0.74$ から $b = -0.83$ まではさらに順列を組合せとする過剰般化が大きくなるため負の転移が起こる．

順列は完全正答で組合せが部分点である生徒の割合を増やす (b が増加する) と，$-0.74 \leq b \leq 0.74$ のとき，a は小さくなり Near overlap singularity 現象が起こる．順列を組合せとする過剰般化が小さくなるため正の転移が起こる．

9.2.2　相互関係の問題を準正答した生徒 (記号計算 : 第 1 回から第 2 回)

組合せの学習直後に，「1 人 1 人の役職を選ぶ」問題を組合せを用いて解答し，準正答した生徒を対象とするため第 1 回から第 2 回考査の変化を分析する．このとき順列は完全正答して組合せは部分点の生徒と，組合せは完全正答して順列は部分点の生徒を比べて 2 つの概念を理解していく過程について考察する．

a 変化のシミュレーション

a を変化させたシミュレーションのダイナミクスと訓練損失の変化と損失曲面上のダイナミクスを図 9.11 に示す．

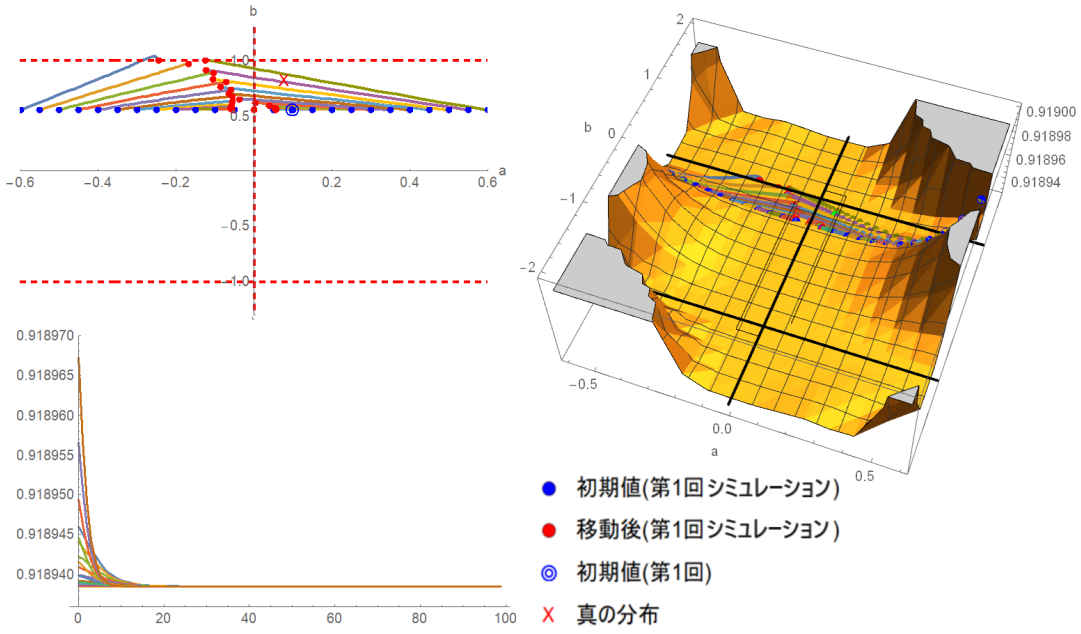


図 9.11　記号:a 変化: 準正答

順列を完全正答して組合せが部分点である生徒の割合の方が大きく偏りがある状態 ($b = 0.55$) から学習を始める．

組合せを順列とする過剰般化が大きくなる (a が減少する) と，臨界直線 $b = 1$ の影響を受けて Elimination singularity 現象が起こる．順列を完全正答して組合せは部分点の生徒のみになり，順列と組合せ共に完全正答した生徒がいなくなる．組合せを順列とする過剰般化が $a < -0.5$ より小さくすると，組合せの学習が進まずに順列の理解の変容が起こる．

順列を組合せとする過剰般化が大きくなる (a が増加する) と，臨界直線 $a = 0$ の影響を受けて $v = 0.25$ より w_1，w_2 が 0.25 で過剰般化の差がなくなる状態 (Overlap singularity 現象) になり，さらに臨界直線 $a = 0$ を超えて停滞し Cross overlap singularity 現象が起こる．順列は完全正答して組合せが部分点である生徒の割合が少し増えるため，組合せの学習が進まずに順列の理解の変容が起こる．

b 変化のシミュレーション

b を変化させたシミュレーションのダイナミクスと訓練損失の変化と損失曲面上のダイナミクスを図 9.12 に示す．

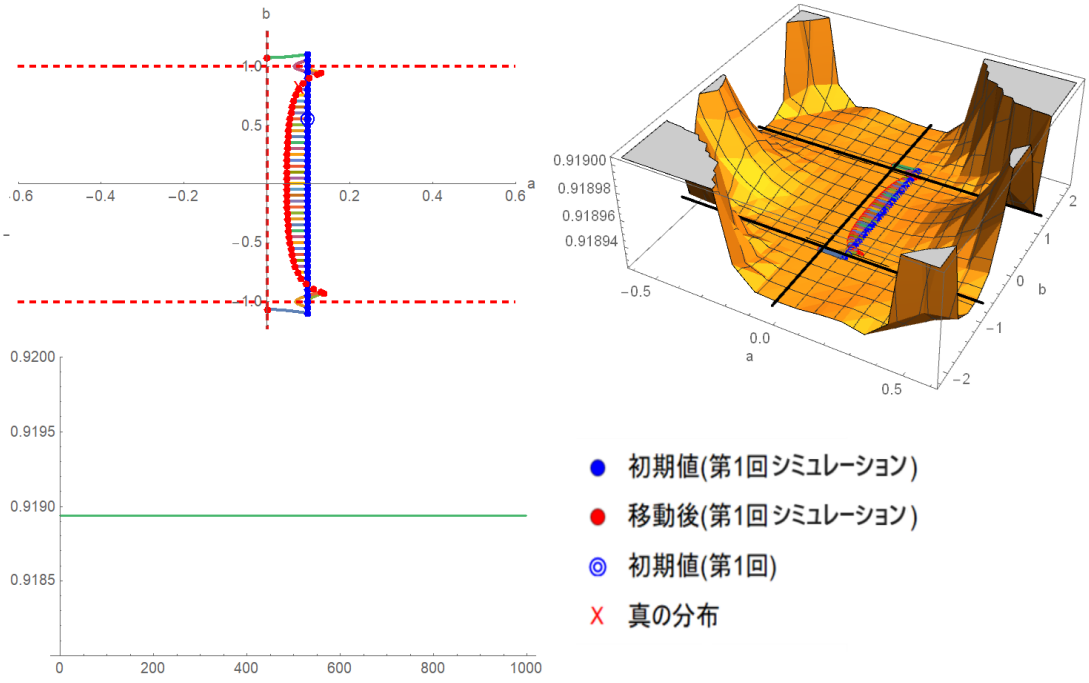


図 9.12　記号:b 変化: 準正答

順列を組合せとする過剰般化が少し大きい状態 ($a = 0.09$) から学習を始める．

順列は完全正答して組合せが部分点の生徒の割合を増やす (b が増加する) と，臨界直線 $a = 0$ の影響を受けて Overlap singularity 現象が起こる．順列を組合せとする過剰般化が大きい状態から，学習の過程で $v = 0.25$ より w_1，w_2 が 0.25 で組合せを順列とする過剰般化との差がない状態 (Overlap singularity 現象) になるため正の転移が起こる．

また，組合せは完全正答して順列が部分点の生徒の割合を増やす (b が減少する) 場合も同様であり，正の転移が起こる．

9.2.3　相互関係の問題を準正答した生徒 (記号計算 : 第 2 回から第 3 回)

組合せの学習後時間が経過した後に，「1 人 1 人の役職を選ぶ」問題において組合せを用いてしまい準正解である生徒を対象として第 2 回から第 3 回考査の変化を分析する．このとき順列は完全正答して組合せは部分点の生徒と組合せは完全正答して順列は部分点の生徒を比べて順列又は組合せが部分点である生徒について考察する．

a 変化のシミュレーション

a を変化させたシミュレーションのダイナミクスと訓練損失の変化と損失曲面上のダイナミクスを図 9.13 に示す．

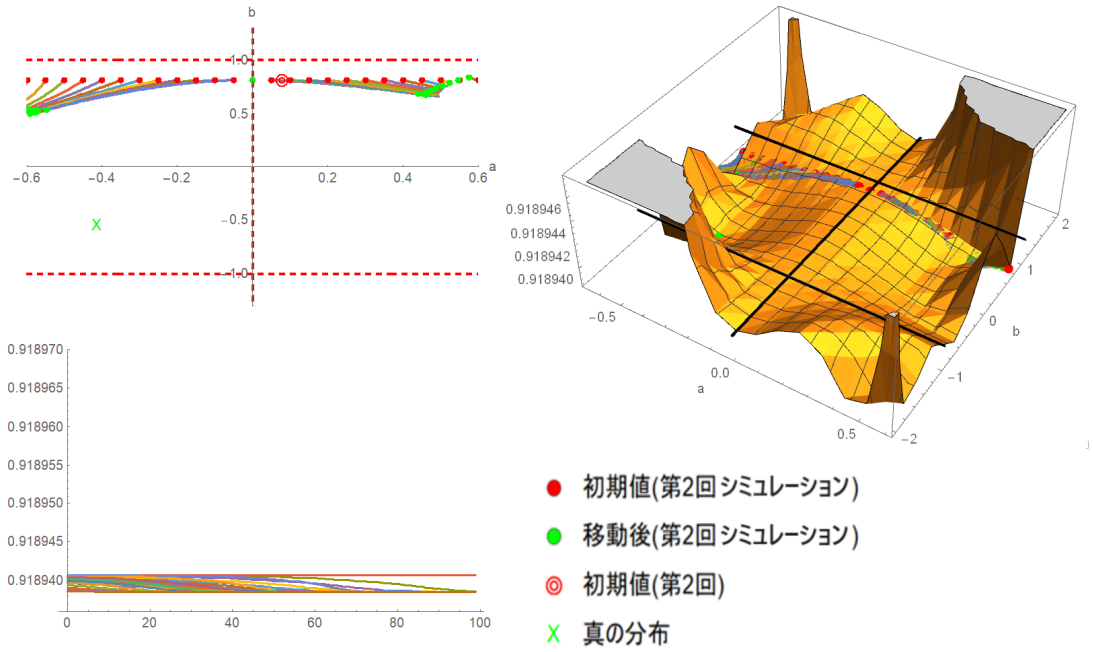


図 9.13　記号:a 変化: 準正答

順列を完全正答して組合せが部分点である生徒の割合の方が多く偏りがある状態 ($b = 0.80$) から学習を始める．

組合せを順列とする過剰般化が大きくなると，臨界直線 $a = 0$ の影響を受けずに $a = -0.57$ で収束 (Fast convergence 現象) する．順列は完全正答して組合せが部分点である生徒の割合が $b = 0.80$ から $b = 0.49$ まで減るため，組合せの理解の変容が起こる．

また，順列を組合せとする過剰般化が大きくなる (a が増加する) と，同様に収束 (Fast convergence 現象) して，組合せは完全正答して順列が部分点である生徒の割合が少し増えるため，組合せの理解の変容が起こる．

b 変化のシミュレーション

b を変化させたシミュレーションのダイナミクスと訓練損失の変化と損失曲面上のダイナミクスを図 9.14 に示す．

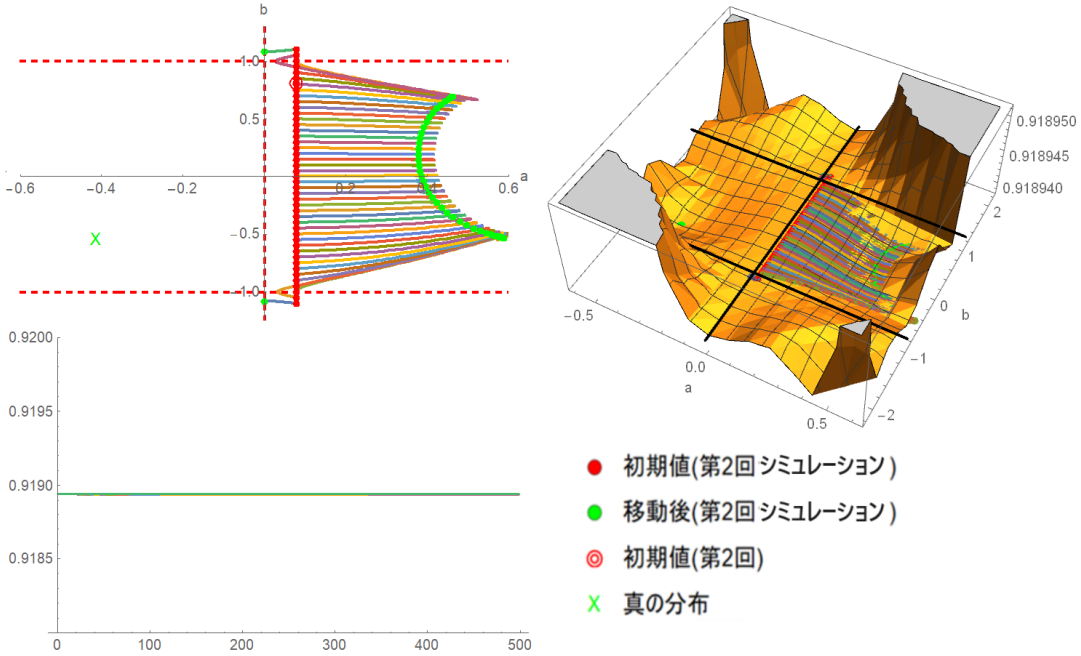


図 9.14　記号:b 変化: 準正答

順列を組合せとする過剰般化が少し大きい状態 ($a = 0.03$) から学習を始める．

組合せは完全正答して順列が部分点である生徒の割合を増やす (b が減少する) と，臨界直線 $a = 0$ の影響を受けずに Fast convergence 現象が起こる．順列を組合せとする過剰般化が初めは小さくなるがその後大きくなり，負の転移が起こる．

また，順列は完全正答して組合せが部分点である生徒の割合を増やす (b が増加する) と，臨界直線 $a = 0$ の影響を受けずに Fast convergence 現象が起こる．順列を組合せとする過剰般化が大きくなり，負の転移が起こる．組合せの学習を進めることで順列を組合せとする過剰般化は $a = 0.46$ から $a = 0.37$ に抑えることができる．

9.2.4　相互関係の問題を正答した生徒 (記号計算 : 第 2 回から第 3 回)

組合せの学習後に時間が経過した後に,「1 人 1 人の役職を選ぶ」問題において，順列の再学習後に相互関係を考慮して正解できる生徒を対象とするために第 2 回から第 3 回考査の変化を分析する．このとき，組合せは完全正答して順列は部分点の生徒と，順列と組合せともに完全正答した生徒を比べて，2 つの概念を理解していく過程について考察する．

a 変化のシミュレーション

a を変化させたシミュレーションのダイナミクスと訓練損失の変化と損失曲面上のダイナミクスを図 9.15 に示す．

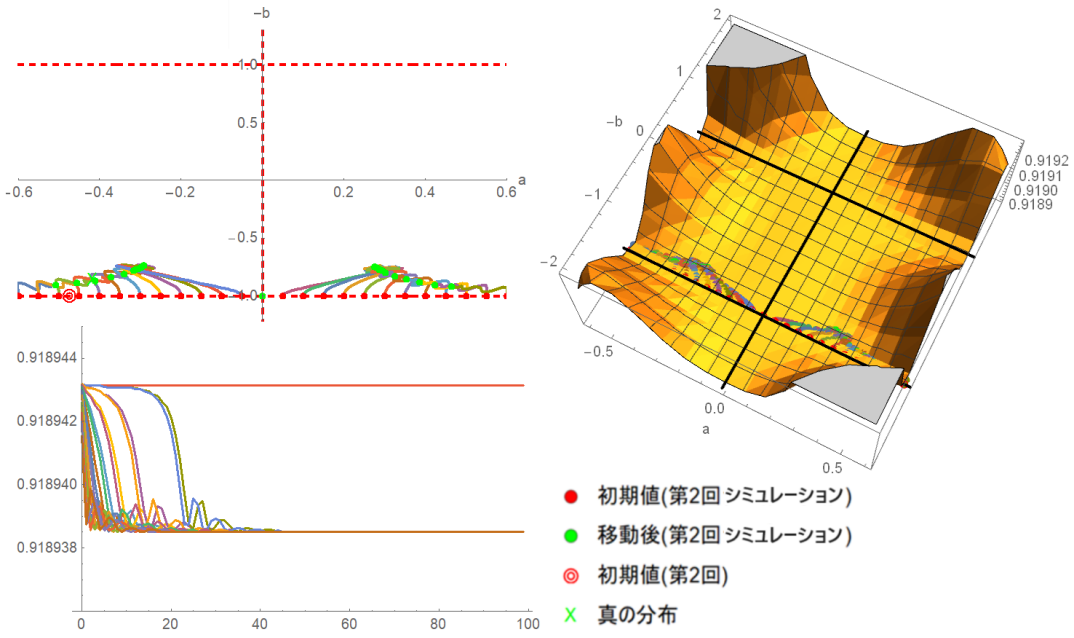


図 9.15　記号:a 変化: 正答

順列と組合せともに完全正答した生徒のみである状態 ($b = 1.0$) から学習を始める．

組合せを順列とする過剰般化が大きくなる ($a > -0.3$ の範囲で減少する) と臨界直線 $a = 0$ の影響を受けずに $a = -0.29$ で収束 (Fast convergence 現象) する．さらに組合せを順列とする過剰般化が大きくなる ($a < -0.3$ の範囲で減少する) と臨界直線 $a = 0$ の影響を受けて停滞して Near overlap singularity 現象が起こる．順列と組合せともに完全正答した生徒の割合が $b = 1.0$ から $b = 0.73$ まで減るため，組合せの理解の変容が起こる．

また，順列を組合せとする過剰般化が大きくなる (a が増加する) と，$a = 0.29$ で収束して同様の状態になり，組合せの理解の変容が起こる．

b 変化のシミュレーション

b を変化させたシミュレーションのダイナミクスと訓練損失の変化と損失曲面上のダイナミクスを図 9.16 に示す．

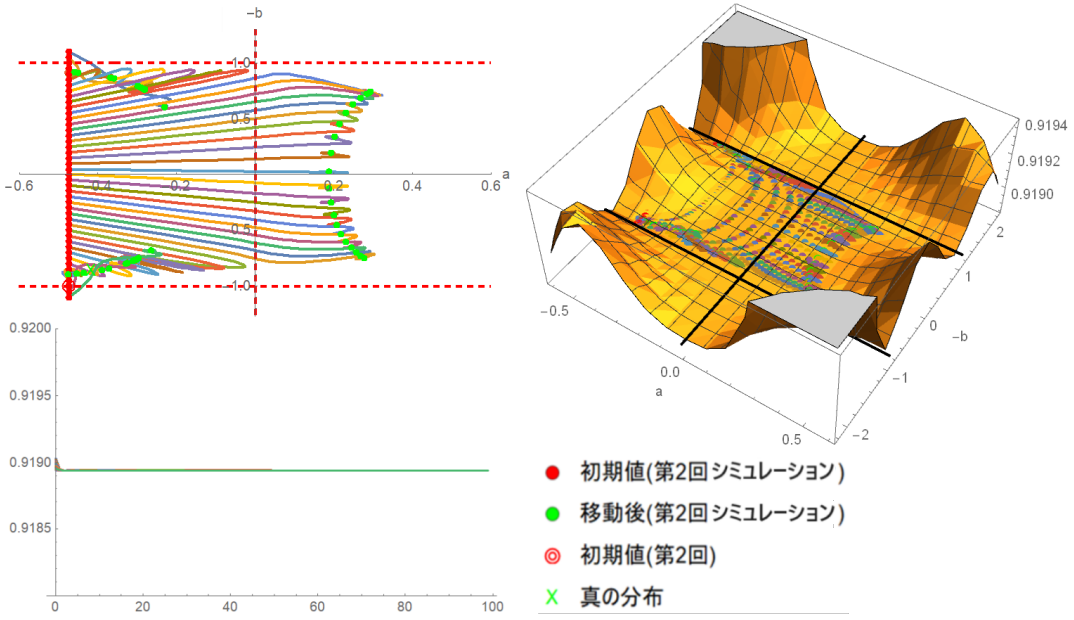


図 9.16　記号:b 変化: 正答

組合せを順列とする過剰般化が大きい状態 ($a = -0.47$) から学習を始める．

組合せは完全正答して順列が部分点の生徒の割合を増やす (b が減少する) と，臨界直線 $b = 1$ に近づき元に戻る Near elimination singularity 現象が起こる．組合せを順列とする大きい過剰般化が抑えらえるため正の転移が起こる．その後，b が減少すると，臨界直線 $a = 0$ を超えて Cross overlap singularity 現象が起こる．

初めは順列を組合せとする過剰般化が大きかったが，学習の過程で $v = 0.47$ より w_1 , w_2 が 0.47 で過剰般化がなくなる状態 (Overlap singularity 現象) になる．さらに順列を組合せとする過剰般化が大きくなるため負の転移が起こる．

第 10 章

まとめと今後の課題

本研究は，学習理論に潜む理論構造を発見し，数理モデルを構築することを目標とした．本研究の意義として 3 つのことがあげられる．

第 1 に，数式処理システムを用いて, 学習モデルによって真の分布を実現できるパラメータの集合は, ヒルベルトの基底定理を適用すると有限個の多項式によって定義される共通零点の集合になることが知られている．第 4 章において数式処理システムを用いた漸化式の計算をさせることで証明を行った (副論文 [25])．また，第 4 章において有限個の多項式で定義された消去イデアルのグレブナ基底を考えることによって, 代数的集合のパラメータ表示を求めた (副論文 [26])．代数的集合の計算 (手法 1) において, 学習モデルが $H = n$ で真の分布が $H_0 = m$ である場合から学習モデルが $H = n - k$ で真の分布が $H_0 = m - 1$ である場合に帰着させた．手法 1 を繰り返すことにより真の分布が $H_0 = 0$ で実現される場合に帰着できた．代数的集合の計算 (手法 2) において学習モデルが $H = n$ で真の分布が $H_0 = 0$ である場合を考察した．本研究では活性化関数として双曲線正接を用いて学習モデルが $H = n$ で真の分布が $H_0 = m$ である場合の代数的集合のパラメータ表示を手法 1 と手法 2 を用いて代数的集合を 3 つの部分 ((1)1 つの中間ユニットに結合する部分, (2)zero 部分, (3)1 次従属な部分) に分けて記述した．

第 2 に，第 5 章において数式処理システムを用いて, 学習モデルとして 3 層パーセプトロンであるニューラルネットワークを作成した．第 6 章において学習モデルの初期値を変えることによってパラメータが臨界直線に対してどのように影響を受けながら動くのかについて考察した．パラメータの変化のグラフを調べ，特異点の近くにおける損失関数の変化と学習損失曲面上の学習のダイナミクスの変化を調べた ([28])．Overlap singularity 現象や Elimination singularity 現象から遠く離れた学習のダイナミクスは未解明である．本研究では, 学習の初期値を連続的に変化させ，遠くから特異領域に近づいたとき，Guo

らの分類に従ったダイナミクスの種類が変化することを調べ，特異点の近くで起きるプラトー現象を明らかにした．

第 3 に，数学の学習において A と B の 2 つの項目の概念が共通概念を含む場合，一方だけではなく，それに関連付けて理解することで深い理解に結びつくことがある．第 7，8，9 章において高等学校数学科の「場合の数と確率」の単元における「順列」と「組合せ」の 2 つの概念を理解する過程と生徒が誤る一因でもある「過剰般化」について取り上げ，それらに関して，ニューラルネットワークを用いた損失曲面により数理モデリングし，過剰般化をもとにした特異領域について，その曲面上でのシミュレーションを行った (副論文 [29], 副論文 [30])．このとき 2 つの生徒集団に対して Overlap singularity 現象を「考査の平均点が等しい」, Elimination singularity 現象を「一方の生徒集団のみになる」状態とした．分析の方法としては，第 7 章において数式処理システムを用いて 3 層のニューラルネットワークを作成し，生徒のテストの平均点の差と，比較する 2 つの生徒集団に対してその人数の割合をもとにした重みを設定して，3 回のテストの理解度をもとに損失曲面を作成した．これらに関して，第 8 章において特異領域が明確になるようにパラメータ変換を行い，座標変換したパラメータ a, b を順列と組合せの平均点の差，2 つの生徒集団の人数の割合と定めた．第 9 章において特異領域を表すパラメータ a, b に対して，作成した損失曲面において初期値を変化させ，シミュレーションを行う方法を確立した．

今後の研究対象として，学習の状況を可視化した際に現れる特異点現象の構造を解き明かしたい．ニューラルネットワークの層の数や中間ユニット数を増やして, パラメータを多く設定することで複雑な高次元の特異点構造が現れる際, 数学の学習における現象をパラメータに対応させることによって数学の学習の状況を捉えることを目標としたい．

また，ニューラルネットワークの構造を改良してより精度の高い学習損失曲面を描画し, 特異領域 Overlap singularity や Elimination singularity における現象以外の曲面上のダイナミクス (損失の変化) が特異点を持つ場合について数学学習の現象を対応させて調べたい．

さらに，数学学習における現象において, 2 つの概念を理解する際の過剰般化現象について高等学校において「順列・組合せ」以外に生徒が数学の概念を深く理解する過程について学習理論を用いて数学的な根拠から解き明かしたい．過剰般化現象については，生徒集団 (i), (ii), (iii) と対応させて生徒の理解が変化する様子を可視化したい．過学習や理解する過程で生徒が躓きやすい箇所で起きる現象についても汎化損失や学習損失の関係を研究したい．そのうえで，多くのサンプル数で複数年度データを取り詳しく分析することで, 理解の過程を明らかにして教師の効果的指導法について提案したい．

謝辞

本論文は筆者が甲南大学大学院自然科学研究科知能情報学専攻博士後期課程に在籍中の研究成果をまとめたものである．

本論文の作成にあたり，多くの方々にご指導ご鞭撻を賜りました．

指導教員である高橋正教授には，現在まで 15 年に渡り研究方法に始まり研究活動全般を長期履修制度によりご指導いただきました．数学を理解することの難しさ奥深さそして創り出すことの楽しさを改めて認識することができたのは高橋先生のおかげです．心からお礼を申し上げ，感謝致します．

甲南大学大学院自然科学研究科知能情報学専攻　森元勘治教授，渡邊栄治教授，小出武教授，田村祐一教授，梅谷智弘教授には，適切なご助言を賜りました．心から感謝申し上げます．

順天堂大学数理・データ科学教育研究センター　大橋真也教授には数式処理や統計的な処理の手法をはじめ様々なご意見ご助言頂きました．平井崇晴先生には数学教育について数多くの助言を頂きました．心から感謝致します．

本論文の教育実践において実際の授業に協力してくれた生徒の皆さんに，心よりお礼を申し上げます．

最後にいつも私を励まし続けて支えてくれた両親に心から感謝致します．

参考文献

[1] 白畑知彦, 若林茂則, 村野井仁: 詳説 第二言語習得研究. 研究社 (2010)

[2] 渡辺澄夫, 萩原克幸, 赤穂昭太郎, 本村陽一, 福水健次, 岡田真人, 青柳美輝: 学習システムの理論と実現. 森北出版株式会社 (2018)

[3] S. Watanabe: Algebraic geometry and statistical learning theory. Cambridge Univesity Press (2009)

[4] S. Watanabe: Algebraic geometrical methods for hierarchical learning machines, Neural Networks, Vol. 14, No. 8, pp. 1049-1060 (2001)

[5] 渡辺澄夫, 福水健次, 萩原克幸, 甘利俊一: 特異モデルの学習理論, 電子情報通信学会論文誌, Vol. J88-D2, No. 2, pp. 159-169 (2005)

[6] S. Watanabe: Mathematical theory of bayesian statistics. CRC Press (2018)

[7] K. Fukumizu: Likelihood ratio of unidentifiable models and multilayer neural networks, The Annals of Statistics, Vol. 31, No. 3, pp. 833-851 (2003)

[8] K. Fukumizu and S. Amari: Local minima and plateaus in hierarchical structures of multilayer perceptrons, Neural Networks, Vol. 13, No. 3, pp. 317-327 (2000)

[9] S. Amari: Information geometry and Its applications. Springer (2016)

[10] 甘利俊一, 尾関智子, 朴慧暎: 神経多様体の特異点と学習, 日本神経回路学会誌, Vol. 10, No. 4, pp. 189-200 (2003)

[11] S. Amari, T. Ozeki, R. Karakida, Y. Yoshida, M. Okada: Dynamics of Learning in MLP: Natural Gradient and Singularity Revisited, Neural Computation, Vol. 30, No. 1, pp. 1-33 (2018)

[12] H. Wei, J. Zhang, F. Cousseau, T. Ozeki, and S. Amari: Dynamics of learning near singularities in layered networks, Neural Computation, Vol. 20, No. 34, pp. 813-843 (2008)

[13] F. Cousseau, T. Ozeki, and S. Amari: Dynamics of learning in multilayer per-

ceptrons near singularities, IEEE Transactions on Neural Networks, Vol. 19, No. 8, pp. 1313-1328 (2008)

[14] F. Cousseau, T. Ozeki, and S. Amari: Dynamics of learning near singularities in radial basis function networks, Neural Networks, Vol. 21, No. 7, pp. 989–1005 (2008)

[15] W. Guo, H. Wei, Y. Ong, J. R. Hervas, J. Zhao, H. Wang, K. Zhang: Numerical Analysis near Singularities in RBF Networks, Vol. 19, No. 1, pp. 1-39 (2018)

[16] 甘利俊一: 深層神経回路網の幾何, 数理科学, No. 689, pp. 46-52 (2020)

[17] 中原忠男: 算数・数学教育における構成的アプローチの研究. 聖文社 (1995)

[18] 鷲野朋広, 高橋正: ニューラルネットワークの入出力計算, 甲南大学紀要 知能情報学編, Vol. 11, No. 2, pp. 371-388 (2018)

[19] T. Nitta: Resolution of singularities introduced by hierarchical structure in deep neural networks, IEEE Transactions on Neural Networks and Learning Systems, Vol. 28, No. 10, pp. 2282-2293 (2017)

[20] 鷲野朋広, 高橋正: ニューラルネットワークにおける過学習に関する分析, 京都大学数理解析研究所講究録, Vol. 2159, pp. 64-74 (2020)

[21] 渡辺澄夫: 代数幾何を用いた問題解決, 数理科学, No. 711, pp. 28-35 (2022)

[22] 梶原健: 代数曲線入門. 日本評論社 (2004)

[23] D. コックス, J. リトル, D. オシー: グレブナー基底と代数多様体入門 (上)(下). 丸善出版 (2023)

[24] 鷲野朋広, 高橋正: 学習モデルにおける補題の証明, 京都大学数理解析研究所講究録, Vol. 2138, pp. 110-118 (2018)

[25] T. Takahashi and T. Washino: On the application of elimination ideal for statistical model, COMPUSOFT, An International Journal of Advanced Computer Technolog, Vol. 9, No. 5, pp. 3685-3689 (2020)

[26] T. Washino and T. Takahashi: Parametrization of statistical models in Three-layer Neural Networks, Proc. of ICIET 2021 19th International Conference on Information and Education Technology Okayama Japan, pp. 386-390 (2021)

[27] 鷲野朋広, 高橋正: 学習モデルにおける特異点構造の分析の分析, 京都大学数理解析研究所講究録, Vol. 2185, pp. 29-46 (2021)

[28] T. Washino and T. Takahashi: On the analysis of singularity structure in learning, Proceedings of the 26th Asian Technology Conference in Mathematics(ATCM 2021), pp. 297-307 (2021)

[29] T. Washino and S. Ohashi: Learning guidance based on the overlap singularity phenomenon, Sci. Math. Japonicae, e-2023, No. 12, pp. 1-20 (2023)

[30] T. Washino and T. Takahashi: Learning guidance based on the elimination singularity phenomenon, Proceedings of 28th Asian Technology Conference in Mathematics(ATCM 2023), pp. 352-361 (2023)

副論文

1. T. Takahashi and T. Washino: On the application of elimination ideal for statistical model, COMPUSOFT, An International Journal of Advanced Computer Technolog, Vol. 9, No. 5, pp. 3685-3689 (2020)
2. T. Washino and T. Takahashi: Parametrization of statistical models in Three-layer Neural Networks, Proc. of ICIET 2021 19th International Conference on Information and Education Technology Okayama Japan, pp. 386-390 (2021)
3. T. Washino and S. Ohashi: Learning guidance based on the overlap singularity phenomenon, Sci. Math. Japonicae, e-2023, No. 12, pp. 1-20 (2023)
4. T. Washino and T. Takahashi: Learning guidance based on the elimination singularity phenomenon, Proceedings of 28th Asian Technology Conference in Mathematics(ATCM 2023), pp. 352-361 (2023)