

XOR問題における誤差曲面の停留点について

On stationary points of error surface in XOR problem

西山 清*, ○落宰 公志*

Kiyoshi Nishiyama*, ○Koushi Ochisai*

*岩手大学工学部情報システム工学科

*Dep. of Computer & Information Science, Iwate University

キーワード: ニューラルネットワーク (neural network), 誤差曲面 (error surface), 停留点 (stationary point),
 H_∞ 学習 (H_∞ -learning), H_2 学習 (H_2 -learning),

連絡先: 〒020-8551 盛岡市上田4-3-5 岩手大学 工学部 情報システム工学科

西山 清, Tel.: 019-621-6475, Fax.: 019-621-6475, E-mail: nishiyama@cis.iwate-u.ac.jp

1. はじめに

ニューラルネットワークにおける学習とは、ネットワークの出力と教師データの誤差に基づいて個々のニューロンにおける結合重みとしきい値を更新する過程である。学習を停滞させる出力2乗誤差曲面の停留点は一般にニューラルネットワークにおいて厄介な問題となっている。

本研究では、XOR問題における2-2-1ネットワークの停留点について10000個の初期重みベクトルを用いて詳細に解析し、その構造を明らかにする。

2. XOR問題とニューラルネットワーク

2.1 排他的論理和 (XOR) 問題

XOR問題とは線形分離不可能なバイナリ問題で、表1に示す4組 ($N_p = 4$) の入出力関係を学習する問題である。本研究ではKuyperらの行った実験²⁾と条件を等しくするため、入力パターン ($z_1^1[p], z_2^1[p]$) に対する所望な出力パターン $z_3^1[p]$ を

$\{(0,0),0.1\}, \{(0,1),0.9\}, \{(1,0),0.9\}, \{(1,1),0.1\}$ と設定した。

Table 1 トレーニングセット
(XORの入出力関係)

p	$z^1[p]$		$z^3[p]$
	$z_1^1[p]$	$z_2^1[p]$	$z_3^1[p]$
1	0	0	0.1
2	0	1	0.9
3	1	0	0.9
4	1	1	0.1

2.2 ニューラルネットワークの構造

本研究で用いるニューラルネットワークは、3層の階層型フィードフォワードネットワークであり、図1のようなそれぞれ入力層2、中間層2、出力層1個のニューロンをそれぞれもつ2-2-1ネットワークを用いる。

このネットワークのニューロンの出力値は次式で表される。

$$\hat{z}_m^3[p] = f \left(\sum_{j=1}^2 w_{j,1}^2 f \left(\sum_{i=1}^2 w_{i,j}^1 z_i^1[p] + \theta_j^1 \right) + \theta_1^2 \right)$$

ただし、 $f(\cdot)$ はシグモイド関数で、 $f(x) = 1/(1 +$

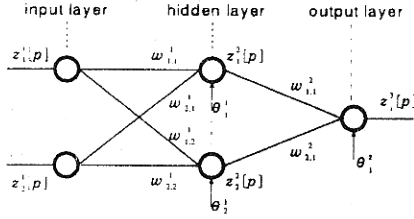


Fig. 1 2-2-1 ネットワーク

$\exp(-\alpha x)$ で表される。 $(\alpha$ は $f(x)$ の傾きであり、 $\alpha=1$ を使用した。) このとき、結合重みとしきい値を要素とする重みベクトル w は次のように 9 次元ベクトルとなる。

$$w = [\theta_1^1, w_{1,1}^1, w_{2,1}^1, \theta_2^1, w_{1,2}^1, w_{2,2}^1, \theta_1^2, w_{1,1}^2, w_{2,1}^2]^T \in \mathcal{R}^{N_w}$$

2.3 学習アルゴリズム

本研究では次の 2 つの学習アルゴリズムを用いて誤差曲面の解析を行う。

[H₂学習 (g-EKF)]

カルマンフィルタとは、ある状態空間モデルのもとで平均 2 乗誤差を最小とする状態ベクトルの最適推定値を逐次推定するアルゴリズムである。H₂学習は、次のニューラルネットワーク全体を表す状態空間モデル

$$w_{k+1} = w_k \quad (\text{状態方程式}) \quad (1)$$

$$y_k = z^3[k] + v_k \quad (\text{観測方程式}) \quad (2)$$

のもとで平均 2 乗誤差 $J_{KF} = E\{(w_k - \hat{w}_k)^T (w_k - \hat{w}_k) | y_0, \dots, y_k\}$ を評価基準とする。H₂学習アルゴリズム (g-EKF) はこの状態空間モデルを線形近似し、カルマンフィルタを適用することによって得られる。

[H_∞学習 (g-EHF)]

H_∞学習は、 $\gamma_f > 0$ が与えられたとき

$$J_{HF} = \sup_{w_0, v} \frac{\sum_{i=0}^k \|e_{f,i}\|^2}{\|w_0 - \hat{w}_0\|_{-1}^2 + \sum_{i=0}^k \|v_i\|^2} < \gamma_f^2$$

を満たすように重みベクトル ($\hat{z}_{k|k} = \mathcal{F}_f(y_0, y_1, \dots, y_k)$) を決定する。H₂学習と同様に式 (1)、(2) の状態空間モデルを線形化し、H_∞フィルタを適用することによって H_∞学習アルゴリズム (g-EHF) が得られる (Fig.2)。

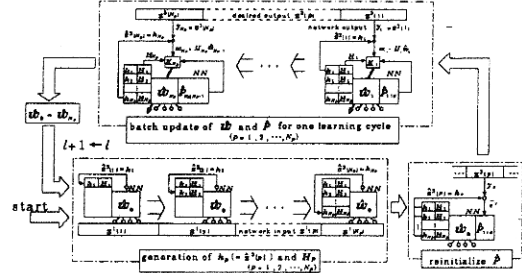
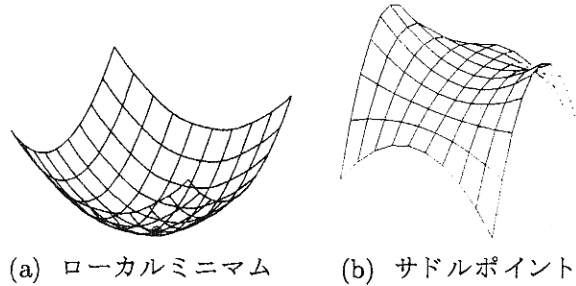


Fig. 2 g-EHF 学習アルゴリズムの動作

2.4 誤差曲面の停留点

本研究における停留点は、誤差曲面において重みベクトルの更新が停滞している領域を指し、学習終了領域を除けば以下のどちらかである。

- ローカルミニマム ... 学習を継続しても誤差がいつまでも減少しない。
- サドルポイント ... 学習を継続するといずれ誤差は減少する。



(a) ローカルミニマム (b) サドルポイント

Fig. 3 誤差曲面における停留点の様子

3. H₂学習からの解析

3.1 重みベクトルの収束領域

学習の打ち切り回数を 10^5 回と設定し、乱数により $(-0.1 \sim 0.1)$ の範囲で初期化された 10^4 種類の重みベクトル $w^{[i]} (i = 1, 2, 3, \dots, 10^4)$ に対し、H₂学習 (g-EKF) を用いて学習を行った。また、学習

は全パターンに対する出力誤差の2乗和

$$J = \sum_{p=1}^{N_p} (z_1^3[p] - \hat{z}_1^3[p])^2 \quad (3)$$

が一定値 (10^{-2}) 以下となるまで繰り返した。その学習結果を表2に示す。これより、停留点でトラップされた試行が43個存在し、また、学習終了時の重みベクトルは、重み投影面上の特定領域に収束していくことが明らかになった。

Table 2 H_2 学習 (g -EKF) による学習結果

最大値	82082
最小値	338
平均値	3273.4
Trap 数	43

このトラップされた43個の重みベクトルには次の特徴があった。

- Case1: ノルムが初期値より小さい。(3試行)
- Case2: 重み $w_{1,1}^1$ と $w_{1,2}^1$ 、あるいは $w_{2,1}^1$ と $w_{2,2}^1$ の値が大きい。(40試行)

以下にそれぞれの場合における重みベクトルの一例をそれぞれ示す。

図4は重みベクトルが特定領域に収束していることを視覚的に示している。(投影面の座標の組合せは重みの収束が見やすい $w_{1,1}^1$ - $w_{2,1}^1$ を用いたが、他の組合せも同様に特定領域に収束していた。)

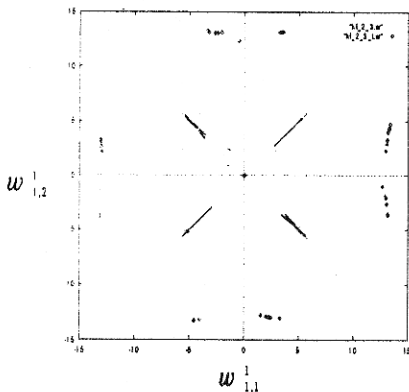


Fig. 4 $w_{1,1}^1$ - $w_{2,1}^1$ における重み投影面 (学習終了点: ドット、停留点: ◇点)

図4からわかるように、トラップされた重みベクトルは投影面上の原点付近 (Case1) と、四方のへり (Case2) で停滞している。

• Case1に分類される重みベクトル (10000試行中3試行)

43試行目; 学習回数: 100000回

重み	初期値	学習打ち切り時
θ_1^1	0.025287745814604	0.049803955418649
$w_{1,1}^1$	-0.007245691335362	-0.020011734607290
$w_{2,1}^1$	-0.033390380108515	-0.027787776462141
θ_2^1	0.011687225680170	0.033087987263041
$w_{1,2}^1$	-0.018221297694785	-0.013943882042974
$w_{2,2}^1$	-0.017045028510623	-0.012153952280338
θ_2^2	0.042122289798561	-0.012861099436199
$w_{1,1}^2$	0.034739499128243	0.015712799157443
$w_{2,1}^2$	0.043347772581621	0.009708851962504
誤差	0.641644326960323	0.639999995510360
ノルム	0.086175269254901	0.074810367357241

• Case2に分類される重みベクトル (10000試行中40試行)

560試行目; 学習回数: 100000回

重み	初期値	学習打ち切り時
θ_1^1	-0.015593411991539	4.000695550007248
$w_{1,1}^1$	0.095139900054249	13.154577542836460
$w_{2,1}^1$	-0.059117339730049	-3.657904140081729
θ_2^1	0.078174997744453	0.306729366519162
$w_{1,2}^1$	0.038083654667768	13.053035710653875
$w_{2,2}^1$	0.055972991554879	2.79493839963534
θ_2^2	0.003316429750401	0.424474859583577
$w_{1,1}^2$	-0.096162807560707	-5.855755622468103
$w_{2,1}^2$	-0.032179412014664	5.431280388834153
誤差	0.640958502069238	0.32000317710134
ノルム	0.183788151324525	21.087557995616116

3.2 誤差曲面の形状

トラップされた2種類の重みベクトルが最終的に誤差曲面上の停留点から脱出できるか考察する。

3.2.1 Case1

本研究の学習判定に用いられている式(3)は、BPの評価基準とも共通している。これを l 回目の学習サイクルにおける出力誤差の2乗和を $J[l]$ とすると、以下のように書き直すことができる。

$$J[l] = \sum_{k=1+lN_p}^{(l+1)N_p} (y_k - h_k(\hat{w}_l))^2 \quad (4)$$

この $J[l]$ からネットワーク全体の重みの偏微分し、符号を反転させることにより、Batch型更新における誤差曲面の最急降下方向が次のように得られる。

$$-\frac{\partial J[l]}{\partial w} = 2 \sum_{k=1+lN_p}^{(l+1)N_p} H_k^T (y_k - h_k(\hat{w}_l)) \quad (5)$$

停留点での誤差曲面の最急降下方向を求めた結果、微小ながらも傾きが存在した。これは更に学習を継続すれば誤差を減少できることを意味している。

実際、学習の打ち切り回数を増加しシミュレーションを行った結果、Case1の全ての試行は図5のような軌跡をとり、12万回以内に学習が終了した。

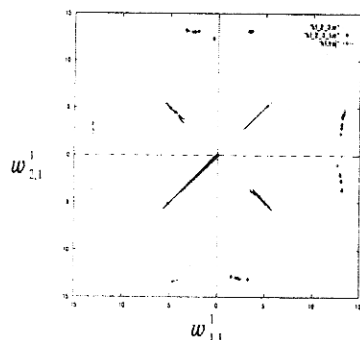


Fig. 5 Case1(43 試行目)における停留点の脱出軌跡

重みベクトルの更新量や最急降下方向、更新重みベクトルの角度などの解析から、Case1の重みベクトルは誤差曲面で図6のような挙動をとったと考えられる。

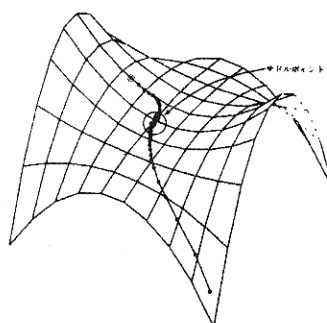


Fig. 6 Case1における誤差曲面の挙動

以上より、Case1の重みベクトルはローカルミニマムではなく、サドルポイントにトラップされていたことがわかる。

3.2.2 Case2

Case1と同様に最急降下方向を調べた結果、投影面のへりでトラップされた重みベクトルにも傾きが存在した。実際に学習の打ち切り回数を 10^8 回まで増やし学習を行った結果、40試行中2試行において学習が終了した。図7にそのときの重みの軌跡を示した。この場合重みベクトルが投影面で大きく折り返している様子がわかる。

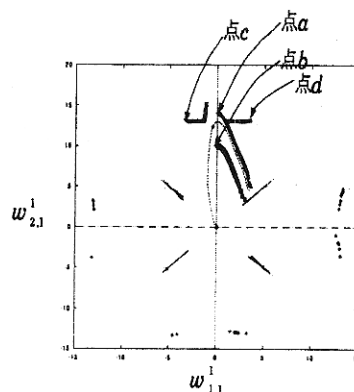


Fig. 7 Case2(2759 試行目)における停留点の脱出軌跡

折り返し点の近傍の形状を解析するために、図7の縦軸(つまり $w_{2,1}^1$)の値が最大値であったときの重みベクトルの値を基点として、基点の周辺の点a、点b、点c、点dを始点として学習を再開した。各点での誤差は表3から基点が最も小さいことがわかる。

Table 3 各点での誤差	
位置	誤差
基点	0.320000618910076
点a	0.320000618914124
点b	0.320000621874026
点c	0.329154021695995
点d	0.459151303406788

この結果と H_2 学習は誤差曲面の『山登り』ができないことを考慮すると、図7のような重みの挙動は、図8のような誤差曲面の形状によるものであると考えられる。ただし、図8の誤差曲面は実際には残りの7次元の重みも考慮しなければならない。

4. H_∞ 学習からの解析

H_∞ 学習(g -EHF)を用いて同様の解析を行うことにする。表4に学習結果を示す。 H_2 学習と比べると、学習終了までに要する回数が大幅に少なく、またトラップした試行も7試行しかなかった。

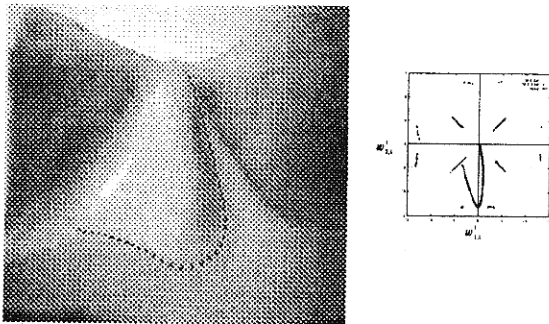


Fig. 8 Case2における誤差曲面と重みベクトルの挙動

Table 4 H_∞ 学習による学習結果; $\gamma_f=1.5$

最大値	74022
最小値	118
平均値	203.0
Trap 数	7

4.1 重みベクトルの収束領域

H_∞ 学習は H_2 学習と同じ初期値から学習を行ったにもかかわらず、学習終了後の重みベクトルの存在領域は大きく異なっていた。ただし、存在領域が原点に対して対称である点は共通していた。

Table 5 (入力層 - 中間層) 間の重みの投影パターン

同一ニューロン間	異なるニューロン間
$w_{1,1}^1 - w_{2,1}^1$	$w_{1,1}^1 - w_{1,2}^1$
$\theta_1^1 - w_{1,1}^1$	$\theta_1^1 - \theta_2^1$

また、前述したように、 H_∞ 学習では学習の打ち切り回数が10万回のとき、7試行しかトラップされなかった。ただし、この7試行は H_2 学習では学習が終了していた。さらに、学習を継続したとき、この7個の重みベクトルが、 $w_{1,1}^1 - w_{2,1}^1$ 投影面でどのような振舞をするか解析した。

Table 6 (中間層 - 出力層) 間の重みの投影パターン

$w_{1,1}^2 - w_{2,1}^2$	$\theta_1^2 - w_{1,1}^2$

Table 7 中間層の前後の重みに関する投影パターン

同一ニューロン間	異なるニューロン間
$w_{1,1}^1 - w_{1,1}^2$	$w_{1,1}^1 - w_{2,1}^2$
$\theta_1^1 - w_{1,1}^2$	$w_{1,1}^1 - \theta_1^2$

その結果、図9のような挙動で学習が終了した。しかし、重みが本来の解の存在領域に収束していないことから、トラップした試行の重みベクトルは異なる領域に収束したと考えられる。これは重み空間の多次元性によるものと考えられる。

4.2 誤差曲面における挙動

H_2 学習において大きく学習を停滞させたCase2の2759試行目に着目し、この試行の重みベクトルが H_∞ 学習を用いた場合、誤差曲面でどのような挙動をとるか解析した。

図10に2759試行目における2つの学習アルゴリズムの学習終了までの軌跡を示す。右図の H_∞ 学習では、停留点でトラップされることなく、わずか184回で学習を終了させた。図11はそのときの誤差曲線、更新量、最急降下方向と更新方向とがなす角度の余弦をそれぞれ示した。これより、 H_∞ 学習は最急降下方向とはほぼ直交するように

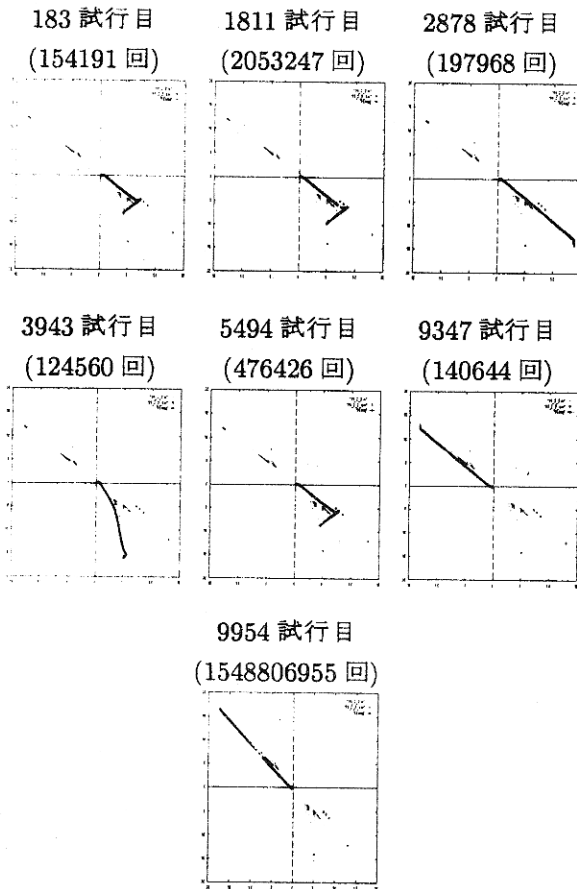


Fig. 9 トラップした試行の $w_{1,1}^1-w_{2,1}^1$ における学習終了までの重みの軌跡 (括弧内は学習回数)

曲面を登り、 H_2 学習とは全く異なった動きをしている。結果として、 H_∞ 学習は図 10 の右の重みベクトルの挙動のように図 8 の溝にトラップされずに、少ない回数で学習を終了している。

5. まとめ

本研究では、XOR 問題における誤差曲面上の停留点について実験的に解析した。Case1 のような重みベクトルは、学習を重ねることで停留点を脱出し、学習を終了した。この結果は Kuyper らの論文にあるような、一定の範囲内の停留点は全てサドルポイントであるという結論²⁾と一致する。しかし、彼らの論文では Case2 のような重み投影面上のへりに存在する停留点は全く考慮していない。

本研究では、この Case2 において学習が終了し

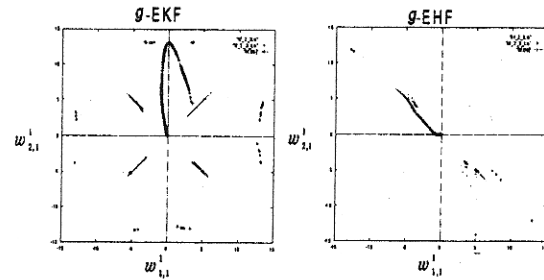


Fig. 10 2つのアルゴリズムによる学習終了までの軌跡

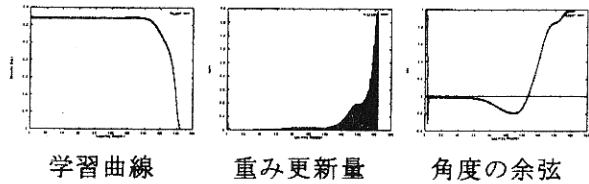


Fig. 11 2759 試行目における H_∞ 学習の学習曲線、重み更新量、角度の余弦

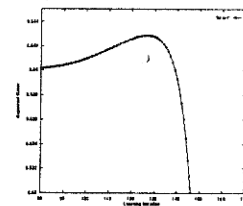


Fig. 12 誤差の上昇

た重みベクトルが誤差曲面上でどのように挙動するかを解析し、誤差曲面の形状を3次元に模擬的に表した。一方で、 H_2 学習では半永久的に学習が終了しないと思われる停留点でトラップされている試行も数多く存在した。このような点でトラップされた場合、 H_∞ 学習でも抜け出すことはできず、この停留点がサドルポイントであるかについてはさらなる解析が必要であると考えられる。

参考文献

- 1) 西山清: 最適フィルタリング, 培風館, 2001.
- 2) I.G.Sprinkhuizen-Kuyper and E.J.W.Boers:
"The error surface of the 2-2-1 XOR network :
The finite stationary points," Neural Networks,
11, pp.683-690, 1998.