ハイパフォーマンス
コンピューティング 86－3
（２００１．５．２５）

# 自然勾配学習法の有効性

田中　研太郎†　　杉原　正顯††　　須田　礼二††

ニューラルネットの学習において、学習の停滞期 (プラトー) が起きて、なかなか学習が進まないことがある。そのプラトーを避け、もっと速く学習する方法として、自然勾配学習法が甘利らによって考えられた[1)]。本論文では、この自然勾配学習法がうまくいかない場合があることを示し、その解決策として、普通の勾配学習法と自然勾配学習法を組み合わせることを提案し、数値実験で有効性を示す。

# Efficiency Of Natural Gradient Learning

TANAKA Kentarou,† SUGIHARA Masaaki†† and SUDA Reiji††

Natural gradient learning (NGL) was proposed by Amari[1)]. In this paper, we show that NGL does not work well in some cases, and introduce combination of ordinal gradient learning (OGL) and NGL to solve the problem.

## 1. はじめに

ニューラルネットの学習方法として、誤差逆伝搬法を用いた勾配学習法がある。この方法は、簡単なアルゴリズムでうまく働くので広く使われるようになった。しかし、この方法で学習していると学習の停滞期 (プラトー) が起きて、なかなか学習が進まないことがある。そのようなプラトーを避けて、もっと速く学習する方法として自然勾配学習法が甘利らによって考え出された[1)]。本論文では、まず第 2 節で自然勾配学習法について説明する。そのあと、第 3 節で自然勾配学習法がうまくいかない場合があることを示し、その場合の解決策を第 4 節で述べる。

## 2. 自然勾配学習法

$n$ 次元の入力$\boldsymbol{x}$を受け取り、それを $m$ 次元の中間層に通し、スカラーの出力 $y$ を出す３層のニューラルネットを考える。入出力関係は式 (1) で与えられるとする。

$$
\begin{aligned}
y(\boldsymbol{x};\boldsymbol{\theta}) &= f(\boldsymbol{x};\boldsymbol{\theta}) + \xi \\
f(\boldsymbol{x};\boldsymbol{\theta}) &= \sum_{\alpha=1}^{m} [v_\alpha \varphi(\boldsymbol{w}_\alpha \cdot \boldsymbol{x} + b_\alpha)] + b_0 + \xi
\end{aligned}
\tag{1}
$$

ここで、$\varphi$ は線形または非線形関数をとる (例:$\varphi(z) = 1/(1+e^{-z})$)。$\xi$ はガウス分布 $N(0,\sigma^2)$ に従うノイズである。今後はニューラルネットのパラメータ $(\boldsymbol{b},\boldsymbol{v},\boldsymbol{w})$ をまとめて $\boldsymbol{\theta}$ で表すとする。それらとは別に、$y^*(\boldsymbol{x})$ という教師の入出力関係があり、入力 $\boldsymbol{x}$ が確率分布 $q(\boldsymbol{x})$ に従って発生するとする。そして、学習データとして、$(\boldsymbol{x}_1, y^*(\boldsymbol{x}_1)), (\boldsymbol{x}_2, y^*(\boldsymbol{x}_2)), \ldots$ が与えられていくとする。これらの学習データをもとにして、ニューラルネットの入出力関係 $y(\boldsymbol{x};\boldsymbol{\theta})$ が、$y^*(\boldsymbol{x})$ になるべく近い入出力関係を持つように $\boldsymbol{\theta}$ を最適化することを学習という。

普通の勾配学習法 (OGL:Ordinary Gradient Learning) では、入力 $x_t$ に対する、ニューラルネットの出力の平均 $f(\boldsymbol{x}_t;\boldsymbol{\theta})$ と、教師の出力 $y_t^*$ との二乗誤差を $e_t(\boldsymbol{x}_t, y_t, \boldsymbol{\theta})$ として、$e_t$ を減らす最急方向 (勾配の逆方向)$-\nabla_\theta e_t$ にパラメータを式 (2) で更新し、最適化を目指す。

$$
\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \nabla_\theta e_t \tag{2}
$$

ここで、$\eta_t$ は各学習のステップにおいて変化してよい数である。式 (2) で使われる二乗誤差のパラメータに対する微分は、高速微分の考え方を使って、その二乗誤差を求める手間の定数倍の手間で、求めることがで

† 名古屋大学工学部　現東大
†† 名古屋大学工学研究科計算理工学専攻
Department of Computational Science and Engineering,Graduate School of Engineering,Nagoya University.

きる（誤差逆伝搬法）。

ところで、式 (2) で表される普通の勾配学習法 (OGL) で学習していると、図 1 の点線のように、学習を続けてもなかなか誤差が減らない学習の停滞期（プラトー）が起こる。このようなプラトーが起こる原因として、パラメータ $\boldsymbol{\theta}$ を動かしてもニューラルネットの入出力関係があまり変わらないような場所を、$\boldsymbol{\theta}$ がうろうろしているという事が考えられる。ニューラルネットの動作があまり変わらないということを、ニューラルネットが「近い」と考えて、そこに距離のようなものを考えることができる。その「距離」を考慮に入れてパラメータを動かせば、プラトーを避けられるだろうと予想できる。情報幾何によると、その距離は計量 $g_{ij}(\boldsymbol{\theta})$ を使って $g_{ij}(\boldsymbol{\theta})d\theta^i d\theta^j$ で表される。計量 $g_{ij}(\boldsymbol{\theta})$ はフィッシャー情報行列と呼ばれる。式 (7) の形をしたニューラルネットの場合のフィッシャー情報行列は、定数倍を除いて式 (3) で与えられる。

$$g_{ij}(\boldsymbol{\theta}) = E_{q(x)}\left[\frac{\partial f(\boldsymbol{x};\boldsymbol{\theta})}{\partial\theta^i}\frac{\partial f(\boldsymbol{x};\boldsymbol{\theta})}{\partial\theta^j}\right] \tag{3}$$

このとき、$e_t$ の最急方向はフィッシャー情報行列 $G(\boldsymbol{\theta})$ を使って $-G^{-1}(\boldsymbol{\theta})\nabla_\theta e_t$ となるので、式 (2) の学習の更新則も以下のように変形される。

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t G_t^{-1}\nabla_\theta e_t \tag{4}$$

式 (4) の方法で学習する方法を自然勾配学習法 (NGL:Natural Gradient Learning) と呼ぶ。

実際にはフィッシャー情報行列を求めるのは難しい。そこで、フィッシャー情報行列を、式 (5) に従って逐次的に推定していく方法が考え出された[2)]。

$$\hat{G}_{t+1}^{-1} \sim (1+\epsilon_t)\hat{G}_t^{-1} - \epsilon_t\hat{G}_t^{-1}\nabla_\theta f_t\left(\nabla_\theta f_t\right)^T\hat{G}_t^{-1} \tag{5}$$

ここで、$\epsilon_t$ はフィッシャー情報行列の推定係数で、各学習のステップにおいて変化してよい適当な数である。最初の $\hat{G}_1$ としては、単位行列を用いることにしている。この方法で推定した $\hat{G}$ を使って、あとは NGL と同じように、式 (4) に従ってパラメータを更新していく学習方法を、ANGL(Adaptive Natural Gradient Learning) と呼ぶ。ANGL は、NGL の近似的な学習方法になっている。

また、学習を続けることにより、パラメータがどのくらい最適化されているのかを、各学習ステップにおいて評価したいとする。そのような誤差の評価尺度としては、汎化誤差（ニューラルネットと教師の出力の二乗誤差を入力について平均したもの）が用いられる。しかし、これを実際に求めるのは難しい。その代わりとして、学習データから計算した訓練誤差 $\frac{1}{T}\sum_{t=1}^{T} e_t(x_t, y_t, \boldsymbol{\theta})$ が評価尺度として使われる。ただし、訓練誤差は、学習の最初の方の大きい誤差が残るため、誤差の挙動が見えにくい面がある。そこで、今回は誤差として、訓練誤差のうちの最近の履歴だけを使って、式 (6) で表されるものを用いている。こちらの方が訓練誤差よりも汎化誤差に近い挙動をする。

$$m_T = \begin{cases} \dfrac{1}{T}\displaystyle\sum_{t=1}^{T} e_t & (T < N) \\ \dfrac{1}{N}\displaystyle\sum_{t=(T-N+1)}^{T} e_t & (T \geq N) \end{cases} \tag{6}$$

## 3. 自然勾配学習法の問題点—簡単なモデルを例にとって

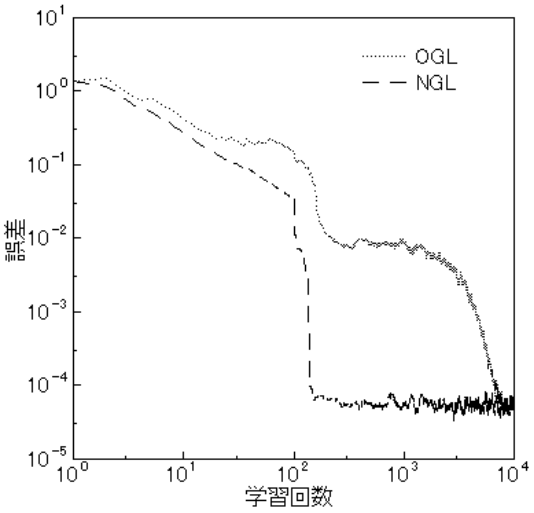


**図 1** 簡単なモデルに対する OGL と NGL の学習曲線

この節では、ニューラルネットとして式 (7) で表されるような簡単なものを考えていく。

$$\begin{aligned} y(x;\boldsymbol{\theta}) &= f(x;\boldsymbol{\theta}) + \xi \\ f(x;\boldsymbol{\theta}) &= v\cdot\exp\left[-\frac{(w\cdot x)^2}{2}\right] \end{aligned} \tag{7}$$

ここで、$\xi$ はガウス分布に従うノイズ $N(0, 10^{-4})$ を表す。このニューラルネットは、スカラー $x$ を入力とし、スカラー $y$ を出力する。ニューラルネットの動作を決定するパラメータは $w, v$ の 2 つだけである。学習の目的の教師の入出力関係は式 (7) と同じ形の以下の式 (8) とする。

$$y^*(x) = \frac{1}{2}\cdot\exp\left[-\frac{(2\cdot x)^2}{2}\right] \tag{8}$$

また、教師の出力にはガウス分布に従うノイズ $N(0, 10^{-4})$ をかぶせる。このときの最適解は、$w = \pm 2, v = \frac{1}{2}$ である。

このような簡単なモデルに対して、OGL,NGL の数

値実験をしたところ図 1 が得られた。学習定数はともに $\eta_t = 0.1$ に設定している。NGL はプラトーを避けて OGL よりも速く学習が進んでいる。

では、いつでも NGL が OGL よりも速く学習するかというとそうでもない。それどころか、自然勾配学習法では、ほとんど学習が進まない、ということが起こりうる。そのことをこれから見ていく。

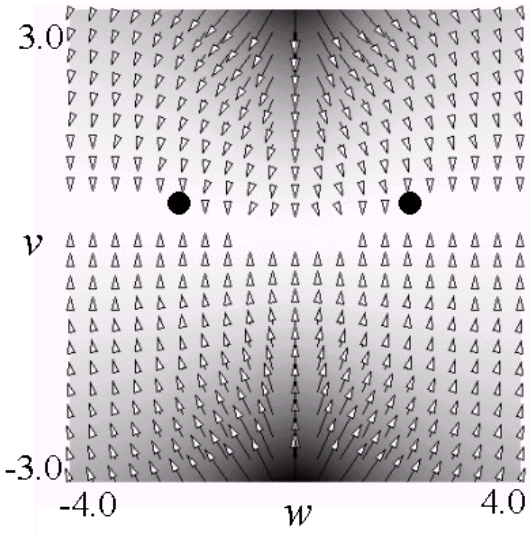


**図 2** OGL の場合の, パラメータ空間の各点における勾配の平均的な向き. 黒丸のところが最適解.

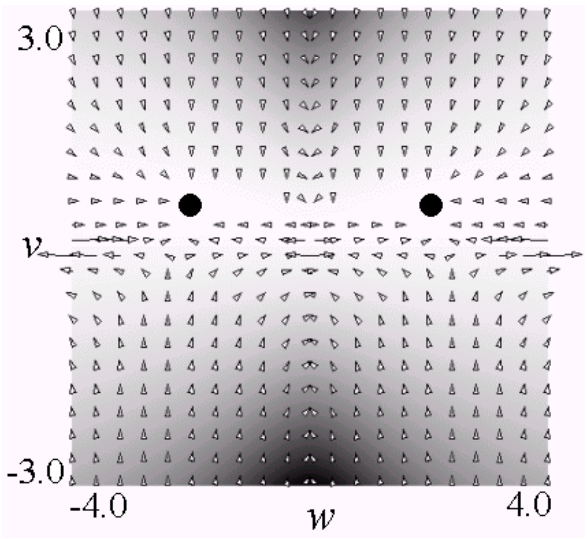


**図 3** NGL の場合の, パラメータ空間の各点における勾配の平均的な向き. 黒丸のところが最適解.

今考えているモデルは、ニューラルネットのパラメータが $w, v$ の 2 つだけなので、そのパラメータ空間の各点において、誤差に対する最急方向が平均的にどこを向くのかを図示できる。OGL,NGL の各々の場合にそれを表したのが図 2,3 である。NGL の図の $v = 0$ の下側では、矢印が奇妙な振る舞いをしていて、最適解（黒丸）に向かいそうもない。そこで、初期値を $v = 0$ の下にとって OGL,NGL で学習させたところ、図 4 のように、NGL では学習が進まなかった。

パラメータ空間のさまざまな点を初期値として学習させて、学習が進まなかった点を黒い点でプロットしたのが、図 5 である。NGL では、$v = 0$ で越えられない境界ができてしまい、初期値が下半面 $(v \leq 0)$ にあると、学習が進みにくくなってしまう。

このことは、直感的には次のように考えられる。$v = 0$ の場合には、$w$ が動いても、式 (7) によりニューラルネットの動作は変わらない。このとき、フィッシャー情報行列は特異になって、その逆行列は発散する。自然勾配学習法は、そのような性質によって、ニューラルネットの出力があまり変わらないような場所では速く動くことができ、プラトーを避けることができるのだが、そのせいで、越えられない境界まででできてしまう。初期値と最適解が境界で隔たっている場合は困ったことになる。

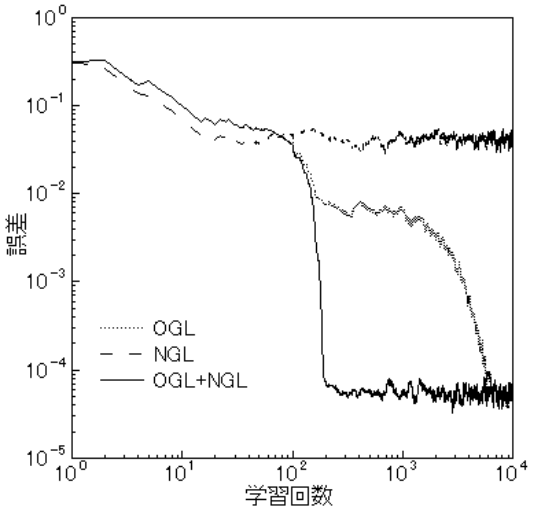


**図 4** 簡単なモデルにおける学習曲線.NGL の場合に学習が進んでいない.OGL+NGL ではちゃんと速く学習が進んでいる.

## 4. 問題点の解決

自然勾配学習法にも、問題点はあり、初期値と最適解が境界で隔たっている場合は困ったことになるという事が、簡単なモデルにおける解析で分かった。そこ

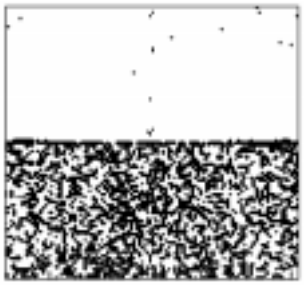

**図 5** パラメータ空間においていろいろな初期値をとった場合に,NGL で学習が進まなかった初期値の点を黒い点でプロットしたもの. 横軸はニューラルネットのパラメータ $w(-4.0 \sim 4.0)$ で, 縦軸は $v(-3.0 \sim 3.0)$ である. 下半面 $(v \leq 0)$ の点を初期値に取ると, 多くの場合で学習が進まない.

で、このような問題点を解決する方法として、NGL では越えられない境界付近では OGL で学習し、OGL でプラトーに陥っているような場合は NGL を使ってその場所から速やかに抜け出す、というような、2つの学習の組み合わせが考えられる。それを実現する手段として、単純な次のようなルールを考えた。

- OGL で学習している場合、各ステップにおける誤差の減少が $\epsilon$ よりも小さい（悪い）ことが K 回続いたら、NGL に切り替える。
- NGL で学習している場合、フィッシャー情報行列の要素の絶対値の最大値が M を超えたとき、OGL に切り替える。

このように OGL と NGL を組み合わせた方法で学習することを、OGL+NGL と書くことにする。

同様に、OGL と ANGL とを組み合わせた OGL+ANGL というものも考えることができる。OGL+ANGL の場合は、上の2つのルールのほかに、次のルールを付け加えるとする。

- OGL から ANGL に切り替わるときは、$\hat{G}$ を単位行列に戻す。

簡単なモデルにおいて、OGL+NGL で学習した結果、図 4 の実線のようになった。学習定数は $\eta_t = 0.1$ と設定した。OGL と NGL の切り替えは以下のようにした。各ステップにおける誤差の減少が、0.002 よりも小さいことが 50 回続いた場合に、OGL から NGL に切り替えた。また、フィッシャー情報行列の絶対値の最大値が、100 よりも大きくなった場合に、NGL から OGL に切り替えた。NGL では学習が進まなかったが、OGL+NGL では、ちゃんと速く学習が進んでいることがわかる。

さらに、図 5 と同じようにパラメータ空間のさまざまな点を初期値として学習させて、学習が進まなかった点を黒い点でプロットしたのが、図 6 である。どの

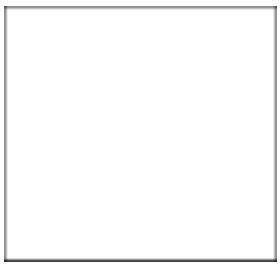

**図 6** 図 5 に対応するものを OGL+NGL でやってみたもの. どの初期値においても学習が進んでいる.

初期値においても学習がちゃんと進んでいる。

次に、もう少し難しいモデルについての学習の数値実験について紹介する。ニューラルネットの入出力関係は、式 (9) で与えられるとする。

$$
\begin{aligned}
y(\boldsymbol{x};\boldsymbol{\theta}) &= f(\boldsymbol{x};\boldsymbol{\theta}) + \xi \\
f(\boldsymbol{x};\boldsymbol{\theta}) &= \varphi\left(\sum_{i=1}^{12} v_i \varphi\left(\sum_{j=1}^{2} w_{ij} x_j + b_j\right) + a_i\right)
\end{aligned}
\tag{9}
$$

ここで、$\varphi(x) = 1/(1+e^{-x})$ である。このニューラルネットは、2つの入力 $x_1, x_2$ を受け取り、スカラー $y$ を出力する。ニューラルネットの動作を決めるパラメータは、$a_i, b_j, w_{ij}, v_i (i = 1 \sim 12, j = 1 \sim 2)$ である。

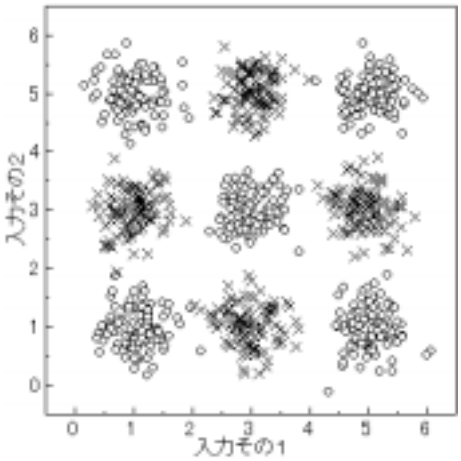


**図 7** 学習させる入出力関係. ○が 1, ×が 0 の出力を表す.

教師の例題は図 7 のように与えた。入力は9個の区画を順番に巡って行き、その中でガウス分布に従って振らせた値を取った。出力は、○が 1 で、×が 0 である。

このモデルでは、フィッシャー情報行列を手計算で求めるのは難しいので、自然勾配学習法として ANGL

を用いた。学習定数は全ての場合に $\eta_t = 0.1$ とし、フィッシャー情報行列の推定係数は、$\epsilon_t = 1/t$ に設定した。また、全てのパラメータの初期値は、$-1$ から 1 の間にとった。そして、OGL+ANGL における OGL と NGL の切り替えは以下のようにした。各ステップにおける誤差の減少が、0.001 よりも小さいことが 100 回続いた場合に、OGL から ANGL に切り替えた。また、フィッシャー情報行列の絶対値の最大値が、100 よりも大きくなった場合に、ANGL から OGL に切り替えた。そのような条件のもとで数値実験したところ、図 8 のようになった。ANGL では学習がうまくいかなかった場合でも、OGL+ANGL では学習がちゃんと速く進んでいる。

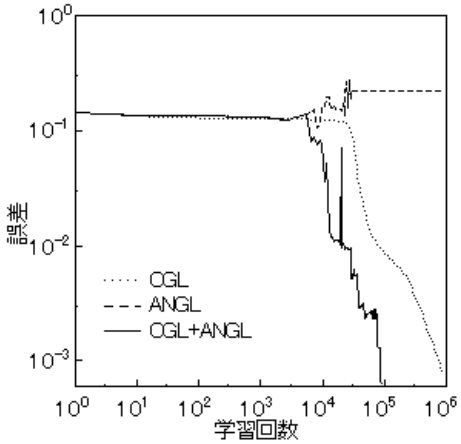


**図 8** 少し難しいモデルにおける学習曲線

## 5. 終わりに

本論文では、自然勾配学習法がうまくいかない場合があることを示し、普通の学習法と自然勾配学習法とを組み合わせた方法を提案した。今後の課題としては、２つの学習法の切り替えの仕方の改良などが考えられる。

## 参 考 文 献


1) S .Amari, Natural Gradient Works Efficiently in Learning , *Neural Computation*, 10, 251-276, 1998.

2) S .Amari, H. Park, and K. Fukumizu, Adaptive method of realizing natural gradient learning for multilayer perceptrons, *Neural Computation*, accepted.