ニューラルネットワークにおける多段階学習法とその学習特性解析

2014年1月

田口　　功

（千葉大学審査学位論文）

ニューラルネットワークにおける多段階学習法とその学習特性解析

2014年1月

田口　　功

概要

機械学習は，1960 年頃から人工知能の一分野として研究が始まり，人工知能研究の一環として，機械が経験から学習することで自動的に行動を獲得していく仕組みを実現することを目指す分野である．現在，その定義は,「データの中で，見えているものから，見えていないものを予測する」技術がほとんどである．

機械学習には，大きく分けて,「教師なし学習 (Unsupervised Learning)」と「教師付き学習 (Supervised Learning)」がある．

教師なし学習では，入力データのみが与えられ，出力すべき値があらかじめ決まっていない．すなわち，与えられたデータを基準なしに自動的に分類する手法である．代表的なモデルは，自己組織化マップ (Self-organizing maps) と言われているもので，コホーネンマップ (Kohonen's Map) とも言われている．この特徴は，色々な要素に対して，いろいろな情報から，似たようなものを分類して仕分けをするのが得意である．分類する要素については，事前に決定しておかなければならない．

それに対して，教師付き学習は，入力データが与えられた時，学習を通して出力を正しく予測することが目的である．ここで，学習中教師データと呼ばれる入力データおよび出力データの事例が多数与えられることで，そこに内在する規則性を学習により獲得できる．獲得後，新たな入力データに対して，正しい出力が出るようにする機械を作ることが目的となる．本論文では，教師あり学習に注目し，画像・音声の処理・認識，制御，時系列解析，予測まで，その応用が多岐にわたる階層型ニューラルネットワークに着目した．様々なニューラルネットワークのモデルが提案されているが，現実的に使用されるのは階層型ニューラルネットワークが非常に多い．階層型ニューラルネットワークに関する基礎は，関数近似・学習・汎化能力であり，それらの性能の向上は，重要となる．層数や素子数を増やせばその入出力関係を複雑にできることは容易に想像できるが，最小の層数，最小の素子数での性能の向上に対する基礎的な研究が重要である．

どのようなデータを用いた学習においても，一般的に入力データと出力データの組み合わせに対し，常に誤差が多く残るデータや逆に比較的誤差が常に少ない入力データと出力データのペアが存在する．このことは，小規模の計算機実験でも容易に確かめることができる．すなわち，学習しにくいデータが存在し，学習するデータを一律には扱うことができない．

そこで，本論文では，階層型ニューラルネットワークを用いた学習において，学習する前に教師データを学習のしやすさに着目した分類，その分類に基づく多段階学習，誤差の大きさに応じた学習係数の動的調整を特徴とした総合的な多段階学習方法を提案した．さらに，出力層素子への入力特性に着目し，学習の継続の可否を判定する振幅減少条件，目標値捕捉条件を提案した．そして，提案手法を検証するために関数近似問題を対象とした計算機実験を行なった．その結果，従来の方法では，学習が困難な対象に対して，提案手法では学習が容易に行われ，学習時間も短縮される結果となった．したがって，従来からシグモイド素子から構成される NN では困難とされている学習を可能とし，その有効性を

示した．また，多段階学習を取り入れることによる学習時間の短縮効果，振動現象の利用に対する有効性も示した．提案手法は特別な素子を用いるわけではなく，一般的なシグモイド素子だけを用いている．したがって，関数近似問題以外の問題に対しても容易に適用可能であると考えられる．また，時系列パターンを用いた予測・推定問題等における学習に対し，学習しにくいデータを多く含む問題にも有効に利用できると考えられる．さらに，本論文では最も基本的な BP 法に基づく効率的学習法を提案したが，BP 法の拡張である QPROP 法や RPROP 法などにも適用可能で，振動特性を積極的に利用し，さらなる誤差の減少を検討するとともに，BP 法の拡張手法への適用を検討した．具体的には, 多段階学習法の学習則として QPROP 法と RPROP 法を用いた場合にも, 学習精度および学習時間の両方が改善されることを計算機実験により確認した．また, 学習においては学習曲線の振動現象が学習性能に重要な役割を果たすと共に, 振動現象の発生メカニズムについて考察した．さらに, 学習が終了しても誤差が小さくならないデータに対して, 学習中の挙動を観察することにより, その理由を検討した．

多段階学習法においては, 学習データを分類する際, 入力ベクトル間の距離に対する出力ベクトル間の距離や出力ベクトルの大きさ (本論文で扱った関数近似問題では出力値) の関係に着目している．今後の課題として, このような距離の関係を利用できない, 例えば判別問題のような学習対象に対しても適用できるように, 多段階学習法をさらに拡張することが挙げられる．

目 次

1 まえがき **1**

2 ニューラルネットワークの歴史および学習曲線の振動 **4**
2.1 ニューラルネットワークの歴史 4
2.2 層構造ニューラルネットワーク 5
2.3 ニューラルネットワーク学習法 6
2.3.1 BP 法 6
2.3.2 QPROP 法 9
2.3.3 RPROP 法 9
2.4 学習曲線の振動 10
2.4.1 ニューラルネットワークにおける学習曲線の振動 10
2.4.2 誤差増加ベクトルおよび誤差減少ベクトル 11

3 多段階学習法 **13**
3.1 誤差に応じた学習係数の値の動的調整 13
3.2 出力層素子への入力特性 13
3.3 教師データ選択方法 14
3.4 効率的多段階学習法 15
3.5 多段階 QPROP 法 18
3.6 多段階 RPROP 法 18

4 関数近似問題を例とした計算機実験 **20**
4.1 関数近似問題を例とした計算機実験 1 20
4.1.1 教師データ 21
4.1.2 教師データの選択 22
4.1.3 比較対象手法と実験の諸設定 24
4.1.4 計算機実験結果 1 24
4.1.5 実験結果 1 の考察 25
4.2 関数近似問題を例とした計算機実験 2 31
4.2.1 教師データ 32
4.2.2 比較対象手法と実験の諸設定 33
4.2.3 計算機実験結果 2 および考察 34
4.2.4 BP 法と振動の有効性 37
4.2.5 振動現象発生のメカニズムと学習中の挙動 37

5 あとがき **46**
参 考 文 献 48 著者関連論文リスト 51 謝辞 53

1　まえがき

近年，入出力がデジタル的なパルスニューラルネットワーク (Pulsed Neural Network,PNN)[1]，複素ニューラルネットワーク[2] やニューロコンピューティングと量子計算機を融合させた量子ニューロコンピューティング[3] が注目され，バックプロパゲーション (Backpropagation, BP) 法を用いたシグモイド素子から構成される階層型ニューラルネットワーク (Neural Network, NN) は古典的なネットワークに成りつつある．しかし，NN は，最近のニューラルネットワークに比較して構成が容易であるために，現実的には BP 法やその拡張手法[4][5] を学習則として取り入れ広範囲に利用されている．学習対象が複雑で，かつ，多量の学習データを用いた学習が必要な場合，学習させても学習が成功しない，あるいは，成功しても多大な学習時間を要するという問題点が残る[15]~[17]．本論文では，この問題を解決するための方法を提案する．

従来から，学習を効果的に進める上での学習係数決定法，教師データの選択法や追加学習，揺らぎ（振動特性）を利用した学習法に関して多く研究されている[6]~[12]．また，慣性係数の利用は，学習曲線の振動を抑え，学習速度の向上に重要な要素であるとされている．

多量の教師データは，学習時間の増加を招くとともに記憶領域を圧迫する．教師データの選択法として，クラス間の最近傍データの組を選択するペアリング法[6] や取り除くと誤差が増加するデータを採用する (Active Data Selection, ADS) 方式が提案されている[7]．また，学習中の誤差を監視して，誤差の大きい教師データのクラス分けを行ない，学習する方法もある[9]．

ペアリング法は，汎化能力を保証するための教師データを必要とする．また，クラス間のデータの距離から教師データクラスを分けることによって，教師データを選択するとともに，クラス数と同数の出力素子数が必要となる．

また，ADS 方式は誤差および近傍教師データからの推定から教師データを選択するシステムが必要となるが，誤差が小さく精度は高い．クラス分けは学習開始後に行なう．したがって，少なくとも学習の初期段階では，誤差が大きいためにすべてのデータを学習することになり，学習中は常に誤差に応じた教師データの調整が必要であるため，ADS 方式は多量のデータを扱う学習において，高精度の学習を可能にするが学習時間の短縮には結びつかず，むしろ長くなる．

本論文では，学習が困難とされている学習対象に対して，学習を可能とし，かつ，学習時間を短縮することを目的とする．学習は多段階とし，学習の各段階で用いる教師データをあらかじめ決定しておく．学習の初期段階ではすべての教師データを用いることはせず，段階的に教師データを追加するため，学習時間が短縮される．学習の初期段階では，学習しにくいと考えられる教師データを用い，最初に学習する．これらがどのようなデータであるかについては ⟨3.3⟩ で述べる．

Cachin は，ネットワークの出力誤差の大きさに応じてデータの提示回数を制御する学習係数決定法を提案している[8]．本論文では，誤差が大きい教師データに対しては学習係数の値を大きく，誤差が小さい教師データに対しては，学習係数の値を小さくするというこ

とを基本とし，学習中，各教師データごとに学習係数の値を動的に調整する方法を取り入れる．この方法は，誤差の小さなデータの提示回数を減らし，逆に誤差の多いデータの提示回数を増やすのと同じ効果が期待できる．

従来から，学習曲線の振動を抑えて学習時間を短縮するために，慣性係数が用いられている．しかしながら学習対象が複雑な場合，振動現象は誤差を減少させるという点から見れば必要であり，強制的に振動現象を抑制する必要はない[13]．複雑な学習対象では，慣性係数を用いると，振動現象は停止するが，用いない場合に比較して誤差は逆に増加するという問題が生じる．また，学習は平均2乗誤差 (以下:RMSE(Root Mean Square Error) に基づいて行われるのが一般的であるために，本論文でも RMSE を使用した．しかし，RMSE 値のみでは，学習の継続，中止の判断には不十分である．文献 (13) においては振動現象に対する明確な判定条件はなく，目視によっていた．本論文では振動現象に関して2つの判定条件を導入し，次の段階の学習に進むか，学習を最初からやり直すかを決定する．したがって，学習をやり直すことになった場合でも，学習時間の増加を最小限に抑えることが可能である．以上のように，提案手法は振動現象の利用，誤差に応じた学習係数の動的調整，および，教師データの選択をしたうえでの多段階学習であるという3つの特徴をもつ．

第1段階として以下の特徴をもつ多段階学習法を提案している[25]．

- 学習データには学習が困難なデータと容易なデータがあることを明らかにし，学習が困難なデータを優先的に学習し，学習が容易なデータを段階的に追加する多段階学習法を提案している．
- 学習率を一定とするのではなく，学習中の各データについて，誤差に応じた学習率を設定している．
- 以上により，学習時間の大幅な短縮を可能とし，高い学習精度を実現している．
- 複雑な学習対象に対しては，学習曲線の振動現象が有効であるといわれており，外部から強制的に振動を与える研究もあるが[12]，多段階学習法は学習中に自発的に振動が生じ，学習誤差減少に重要な役割を果たす．

しかしながら，以下の問題点がある．

(1) 学習法に単純な BP 法（Backpropagation Method）[5],[13] [26]~[27] を用い，他の学習法を組み込むことも可能であるという立場をとっていたが，組み込んだ場合の具体的な性能評価を行なっていなかった．

(2) 学習精度に大きく影響する振動現象[11]がどのようなメカニズムで生じるかが明らかにされていない．

(3) 学習終了時に，他のデータと比較して誤差の大きい特定の学習データが生じるが，その原因を明らかにしていない．

そこで第2段階としては，以下の内容について検討を行なった．

(1') 多段階学習法の学習則に QPROP 法 (Quick Backpropagation Method) [5] と RPROP 法 (Resilient Backpropagation Method)[26] を組み込んだ場合の学習精度と学習時間を比較評価する．

(2') 振動現象がどのようなメカニズムで生じるのかを学習時の重み更新ベクトルを用いて解析する．

(3') 学習終了時に，誤差の大きい特定の学習データに関して，その原因を振動現象と関連させて考察する．

多段階学習（本論文での説明は3段階とする）においては，一括更新で重み係数の更新が行われる．一般的に予測問題等では，重みの更新は，逐次で行われるが，本論文においては，関数学習を実験では行っているために，すべて一括更新であることに注意する．

(1')~(3') の結果，関数学習に対して，多段階学習法に BP 法，QPROP 法，RPROP 法を組み込んだ場合，学習誤差は小さく，精度が改善された．

学習時間について，多段階学習法を組み込むと，文献 (1) の s 段階学習における推定式 $\frac{s+1}{2s}$ に対して，ほぼ合致し，改善されている．

振動現象については，重み更新ベクトルを絶対値誤差が減少するグループと増加するグループに分けた．減少するグループのベクトルが常に多数派に属し，継続する場合には振動は生じない．また，振動の発生については，誤差が減少する方向に重み更新ベクトルが更新されていたとしても，学習の最後までその状態が続くとは一般的に考えられない．特に，誤差が多く残るような教師データが多く存在する場合の関数学習においては，あるエポックで，任意の増加ベクトルが生じたとする．支配的なグループは，減少ベクトルであるために，任意の増加ベクトルの誤差は，さらに増加する．その結果，任意の増加ベクトルの全更新ベクトルに対する割合が大きくなるために，任意の増加ベクトルに属した学習データは，多数派となり，誤差は減少する．学習が進むと，また，誤差の多く残る任意の増加ベクトルが生じる．この繰り返しで振動が生じる．

多段階学習に対する誤差と振動について，学習パターンの中には，学習終了後にも誤差が大きいままのパターンが存在することが大きな特徴で, 最終段階では，必ずすべてのデータが使用される．そのようなパターンは学習方法や学習ごとに変わるのではなく，特定のパターンに固定される．教師データの絶対値が大きいデータは，第1段階では学習しにくいと判断されたデータだけを学習するので, ある程度誤差は小さくなるが, 第2, 第3段階で新たな学習データが加わると, それらデータの初期の誤差は大きいので, 重み修正ベクトルは新たに加わったデータの誤差を減少させる方向に引き摺られる．したがって, それらは少数派に属することになり, 一度減少した誤差が大きくなる．

それに対して, 第3段階で加わった学習データの誤差の絶対値の初期値（乱数により決定）は小さく，5エポック程度の学習により, 小さな範囲で振動するようになり誤差の大小に関する相対的な関係が固定され，この相対的な関係が維持され，少しずつ全体の $RMSE$ 値が減少する．

本論文の構成は以下のとおりである．2. ではニューラルネットワークの歴史と学習法，これまでの研究の振動の扱いについて述べ，3. では多段階学習法，多段階 QPROP 法，および，多段階 RPROP 法について述べる．また 4. においては，関数近似問題を例とした計算機実験を行ない，その有効性を示すとともに，最後に振動現象発生機構の考察を行なう．5. はまとめである．

2　ニューラルネットワークの歴史および学習曲線の振動

2.1　ニューラルネットワークの歴史

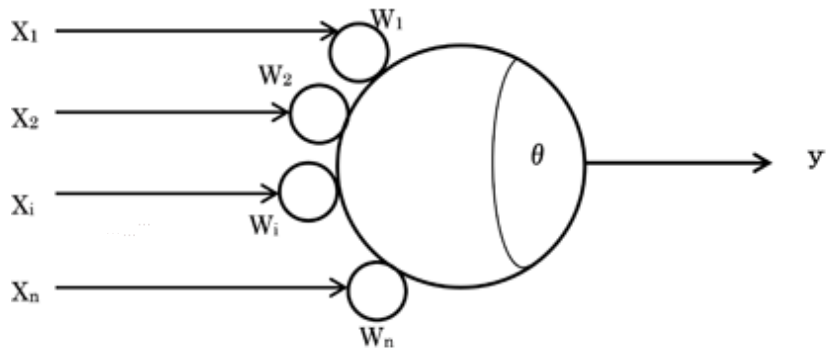

図 1: Neuron Model

1943 年，非常に簡略化された神経回路のモデルを提案したマッカロックとピッツ (McCulloch & Pitts) は，ニューラルネットワークの最初の研究者としてよく知られている．多入力 1 出力素子としてニューロンをモデル化している．そのモデルを図 1 に示す．ここで，x_i は，i 番目の入力，w_i は，シナプス結合の大きさを表わす重み係数であり，結合の強さを示し，y は，出力である．θ は，このニューロンが持つしきい値であり，ニューロンへ伝わる信号がこれを超えると興奮し，以下の場合は興奮しない．式的には，

$$u = \sum_{i=1}^{n} W_i x_i - \theta \tag{1}$$

$$y = f(u) \tag{2}$$

となる．すなわち，式 (1) は，他のニューロンからの重み係数と入力の積和からしきい値を引いた値が，出力関数（式 (2)）の入力 u となる．出力関数のとり得る値の基本として 2 値のみ（0 または 1）を出力する関数の時，これをパーセプトロンという．

$$y = f(u) = \begin{cases} 0 \ (u \leq 0) \\ 1 \ (u > 0) \end{cases}$$

パーセプトロンは，ニューロコンピュータ研究の原点でもあり，これはローゼンブラット（Rosenblatt）により提案された（1958 年）．パーセプトロンは，学習能力を持つ機械と

して注目され，ニューラルネットワーク研究の最初のブームを引き起こした．しかし，この最初のブームは，1969 年に人工知能の父と呼ばれるミンスキーらによってパーセプトロンの限界が指摘され，終わりをとげた．これは，単純パーセプトロン 1 個では，XOR 関数と呼ばれる論理関数を線形分離不可能ということで，実現できない．1 個のニューロンによって実現できる論理関数とできない論理関数あることがわかった．線形分離可能ならば，学習可能となる．1 個のニューロンによって学習できない論理関数を学習するためには複数のニューロンを結合し，ネットワークを構成し，学習すれば学習も可能となる．

60 年代にかけて第一次の研究ブームがあったが，その後 70 年代には，多くの研究者の興味がニューラルネットワークから離れてしまった．

その後のニューラルネットワークの研究は，一部の研究者によって地道に研究が行なわれた．甘利 [34][35]（理論解析），福島 [36]（生理学に基づくニューラルネットワーク），中野 [37]（アソシアトロン），コホーネン [38]（連想記憶モデルと自己組織化）は，今も評価が高い．

1980 年代に入ると再びニューラルネットワークの研究のブームがやってきた．これには，階層構造ニューラルネットワークの比較的簡単な学習法，すなわち，誤差逆伝搬法 [5][30]（Backpropagation ）が提案され，人々の興味を引く例題によって有効性が広く宣伝されたことや，技術的な背景として，コンピュータ技術の進歩があり，ニューラルネットワークのシミュレーションをワークステーションやパソコンで手軽に行えるようになったことが考えられる．

2.2 層構造ニューラルネットワーク

ニューラルネットワークを構造に注目し分類すると, 図 2 に示すように，階層型ニューラルネットワークとニューロン間に結合がある相互結合型ニューラルネットワークに分類できる．

階層構造のニューラルネットワークは，入力ユニットから構成されている層を入力層，出力ユニットから構成されている層を出力層という．その中間にあるユニットは，隠れユニットと呼ばれ，その層は，隠れ層，または，中間層とも言われる．一般的には入力データは，中間層を通って出力層に流れていく．本論文では，階層構造のニューラルネットワークを用い，中間層数は多く設定することもできるが，その数を常に１とした．

相互結合型ニューラルネットワークは，情報の流れとしては，双方向に流れる．すなわち，あるニューロンがほかのニューロンに出力を伝達すると，そのニューロンからも出力を与えられる．

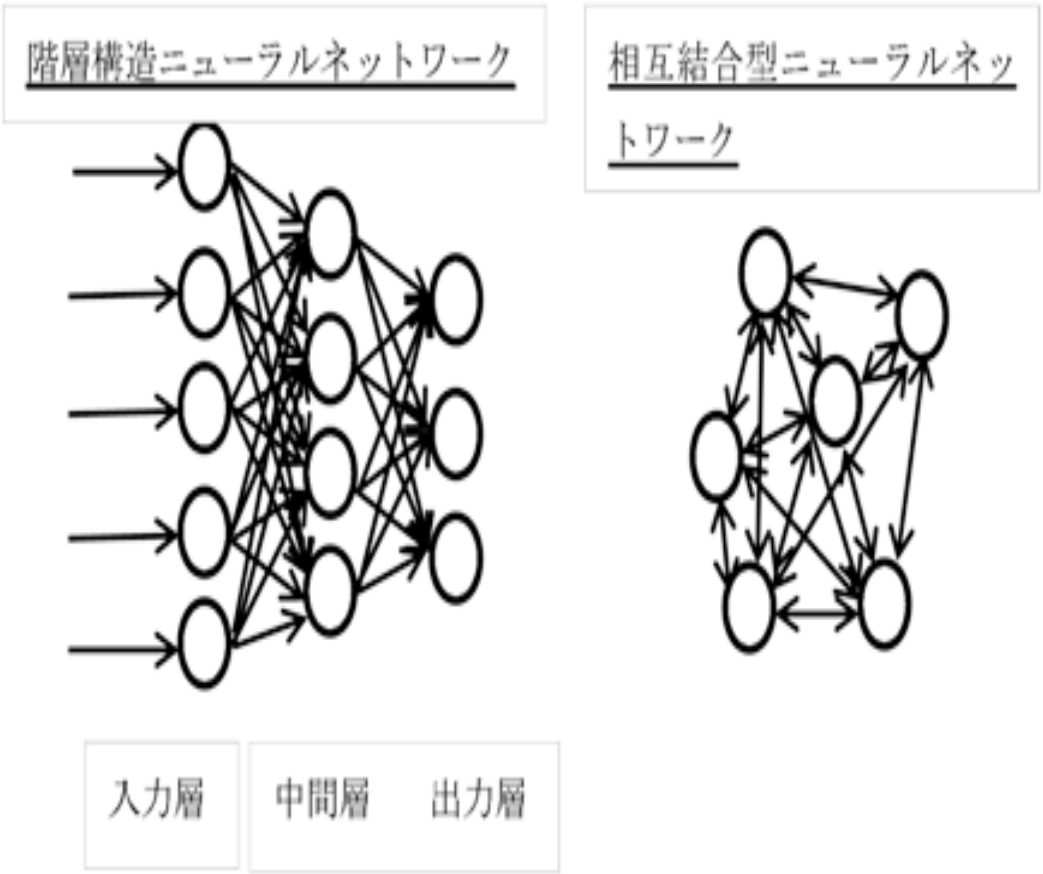

図 2: Classification of Neural Networks by Fundamental structure

2.3 ニューラルネットワーク学習法

2.3.1 BP 法

最初に，図 3 に示すような 3 層ニューラルネットワークを考える．ここで，$\boldsymbol{x}^{(p)}$ は p 番目の入力パターンベクトル，$\boldsymbol{d}^{(p)}$ は，それに対応する教師信号である．$\boldsymbol{o}^{(p)}$ は，実際の出力とする．$V_{ji}(l)$ は，入力層と中間層の間の重み係数，$W_{kj}^{(h)}$ は，中間層と出力層の間の重み係数である．ここで，i は入力数，j は中間層ニューロン数，k は出力層の出力数とする．出力層にある素子は，入力 u に対して，シグモイド関数 $\mathrm{sig}(u)$ を使用すると，

$$\mathrm{sig}(u) = \frac{1}{(1 + \exp(-ku))} \tag{3}$$

となる．学習は，

$$E = \sum_{p=1}^{P} E^{(p)} = \sum_{p=1}^{P} \sum_{k=1}^{K} \frac{(d_k^{(p)} - o_k^{(p)})^2}{2} \tag{4}$$

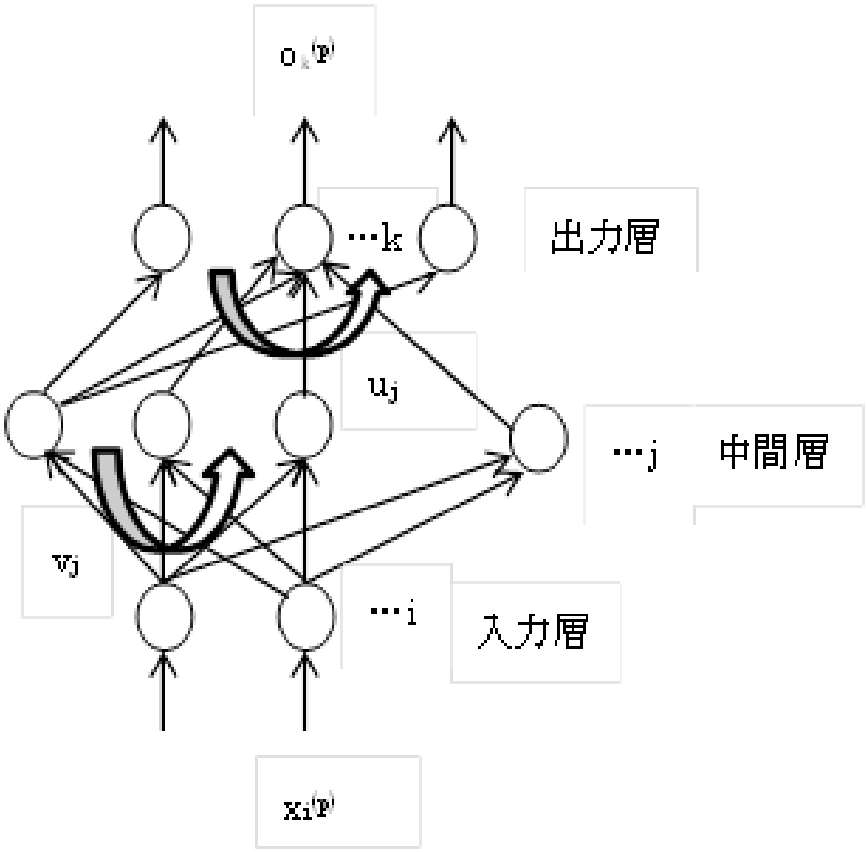

図 3: Three Layered Neural Networks.

の総誤差 E を最小化することである．ここで $d_k^{(p)}$ は p 番目のパターンに対する教師信号であり，$d_k^{(p)} - o_k^{(p)}$ は $\boldsymbol{x}^{(p)}$ に対する 3 層ニューラルネットワークの出力と教師信号の間の誤差となる．パターン $\boldsymbol{x}^{(p)}$ が入力された場合，出力誤差 $E^{(p)}$ は，

$$E^{(p)} = \sum_{k=1}^{m} \frac{(d_k^{(p)} - o_k^{(p)})^2}{2} \tag{5}$$

となる．この $E^{(p)}$ を減少させるように $W_{kj}^{(h)}$ と $\mathrm{V}_{ji}^{(l)}$ を調節する．すなわち，$\Delta W_{kj}^{(h)}$ および $\Delta V_{ji}^{(l)}$ は，それぞれ

$$\Delta W_{kj}^{(h)} = -\epsilon \frac{\partial E^{(p)}}{\partial W_{kj}^{(h)}} \tag{6}$$

$$\Delta V_{ji}^{(l)} = -\epsilon \frac{\partial E^{(p)}}{\partial V_{ji}^{(l)}} \tag{7}$$

となる．ここで，ϵ は学習係数である．出力層にある素子 k の入力 u_k は，$u_k = \sum_{j=1}^{m} W_{kj}^{(h)} y_j^{(p)}$ となり，出力は，$o_k^{(p)} = \mathrm{sig}(u_k)$ となる．また，偏微分の連鎖律を考慮すると，パターン

$p(= 1, \cdots, P)$ に対して，

$$
\begin{aligned}
-\frac{\partial E^{(p)}}{\partial W_{kj}^{(h)}} &= -\frac{\partial E^{(p)}}{\partial u_k}\frac{\partial u_k}{\partial W_{kj}^{(h)}} \\
&= -\frac{\partial E^{(p)}}{\partial o_k^{(p)}}\frac{\partial o_k^{(p)}}{\partial u_k}\frac{\partial u_k}{\partial W_{kj}^{(h)}} \\
&= (d_k^{(p)} - o_k^{(p)})o_k^{(p)}(1 - o_k^{(p)})y_j^{(p)} \qquad (8)
\end{aligned}
$$

が成り立つ．ここで，$\delta_k^{(h)} = (d_k^{(p)} - o_k^{(p)})o_k^{(p)}(1 - o_k^{(p)})$ とおくと，

$$
\Delta W_{kj}^{(h)} = -\epsilon\delta_k^{(h)}y_j^{(p)} \qquad (9)
$$

は，出力層と中間層間の重み係数更新式となる．次に，入力層と中間層との間の重み係数更新も同様に考える．

中間層にある素子 j の入力 v_j は，$v_j = \sum_{i=1}^{l} V_{ji}^{(l)}x_i^{(p)}$ となり，出力は，$y_j^{(p)} = \mathrm{sig}(v_j)$ となる．また，偏微分の連鎖律を考慮すると，パターン $p\,(= 1, \cdots, P)$ に対して，

$$
\begin{aligned}
-\frac{\partial E^{(p)}}{\partial V_{ij}^{(l)}} &= -\frac{\partial E^{(p)}}{\partial v_j}\frac{\partial v_j}{\partial V_{ji}^{(l)}} \\
&= -\frac{\partial E^{(p)}}{\partial v_j}x_i^{(p)} \qquad (10)
\end{aligned}
$$

となる．ここで，$\delta_j^{(l)} = -\frac{\partial E^{(p)}}{\partial v_j}$ とおくと，

$$
\begin{aligned}
\delta_j^{(l)} &= -\frac{\partial E^{(p)}}{\partial y_j^{(p)}}\frac{\partial y_j^{(p)}}{\partial v_j} \\
&= -\frac{\partial E^{(p)}}{\partial y_j^{(p)}}y_j^{(p)}(1 - y_j^{(p)}) \qquad (11)
\end{aligned}
$$

となり，さらに $\frac{\partial E^{(p)}}{\partial y_j^{(p)}}$ は，

$$
\begin{aligned}
\frac{\partial E^{(p)}}{\partial y_j^{(p)}} &= \sum_{k=1}^{n}\frac{\partial E^{(p)}}{\partial u_k}\frac{\partial u_k}{\partial y_j^{(p)}} \\
&= \sum_{k=1}^{n}\delta_k^{(h)}W_{kj}^{(h)} \qquad (12)
\end{aligned}
$$

となる．ここで，

$$
\delta_j^{(l)} = (\sum_{k=1}^{n}\delta_k^{(h)}W_{kj}^{(h)})y_j^{(p)}(1 - y_j^{(p)}) \qquad (13)
$$

とおき，まとめると，

$$
\begin{aligned}
\Delta V_{ji}^{(l)} &= \epsilon\delta_j^{(l)}x_i^{(p)} \\
&= \epsilon(\sum_{k=1}^{n}\delta_k^{(h)}W_{kj}^{(h)})y_j^{(p)}(1 - y_j^{(p)})x_i^{(p)} \qquad (14)
\end{aligned}
$$

となる．

2.3.2 QPROP 法

QPROP 法 [5] および次節で述べる RPROP 法 [26] は，ともに誤差逆伝搬法を改良し，学習の高速化を図った方法である．

QPROP 法は，学習係数をできるだけ大きくしながら，かつ，重み係数の変化量に慣性項を加え，振動を抑えることにより学習の高速化を行なう BP 法の一つである．その慣性項は，現在の重み更新量だけでなく前回の重み更新量を考慮した重み更新を行なう方法である．すなわち，時刻 k における重み修正量を求める時，時刻 $(k-1)$ の重みの修正量を考慮し，

$$\Delta W_{ji}(k) = -\rho^{(k)} S_{ji}(k) + \alpha_{ji}^{(k)} \Delta W_{ji}(k-1) \tag{15}$$

ただし，

$$S_{ji}(k) = \frac{\partial \boldsymbol{E}(k)}{\partial W_{ji}(k)} + \lambda W_{ji}(k) \tag{16}$$

により重みの更新を行なう $(k = 1, 2, \cdots)$．ここで，添字 i は中間層ユニット i，添字 j は出力層ユニット j を表わす．λ は重み変更抑制係数である．学習係数 $\rho^{(k)}$ は，

$$\Delta W_{ji}(k-1) = 0 \ \cup S_{ji}(k) W_{ji}(k-1) > 0 \ \Rightarrow \ \rho^{(k)} = \rho \tag{17}$$

$$\Delta W_{ji}(k-1) \neq 0 \ \cup S_{ji}(k) W_{ji}(k-1) \leq 0 \ \Rightarrow \ \rho^{(k)} = 0 \tag{18}$$

に従って決定する．

また，慣性項 $\alpha_{ji}^{(k)}$ に対しても

$$\tilde{\alpha}_{ji}^{(k)} = \frac{S_{ji}(k)}{S_{ji}(k) - S_{ji}(k-1)} \tag{19}$$

を計算し，

$$\tilde{\alpha}_{ji}^{(k)} > \mu \ \cup S_{ji}(k) \tilde{\alpha}_{ji}^{(k)} \Delta W_{ji}(k-1) < 0 \Rightarrow \ \alpha_{ji}^{(k)} = \mu \tag{20}$$

$$\tilde{\alpha}_{ji}^{(k)} \leq \mu \ \cup S_{ji}(k) \tilde{\alpha}_{ji}^{(k)} \Delta W_{ji}(k-1) \geq 0 \Rightarrow \ \alpha_{ji}^{(k)} = \tilde{\alpha}_{ji}^{(k)} \tag{21}$$

として決定する．すなわち，慣性項に対して $S_{ji}(k) \cong S_{ji}(k-1)$ の場合は，重みの更新量は異常に大きくなってしまうので最大値 μ を設けている．

2.3.3 RPROP 法

RPROP 法は，基本的には，振動を抑えるために前回の重み更新量を記憶しておくことと，前回の重み更新量の符号と今回の重み更新係数の符号を考慮し，５個のパラメーターを条件に応じて調整することにより，振動を抑えた学習を行なう方法である．

$$m = \frac{\partial \boldsymbol{E}(k-1)}{\partial W_{ji}(k-1)} * \frac{\partial \boldsymbol{E}(k)}{\partial W_{ji}(k)} \tag{22}$$

とし，m の符号によって重み更新量が異なる．基本的には，次に示す 3 式を元に重みの更新を行なう．また，更新量が大きすぎないように $\Delta_{\max}$，小さすぎないように $\Delta_{\min}$ を設定していることが特徴である．

$$m > 0 \text{ の時} \begin{cases} \Delta_{ji}(k) = \min(\Delta_{ji}(k-1) * \mu^{+}, \Delta_{\max}) \\ \Delta W_{ji}(k) = -\mathrm{sign}(\frac{\partial \boldsymbol{E}(k)}{\partial \mathrm{W}_{ji}(k)}) * \Delta_{ji}(k) \\ W_{ji}(k+1) = W_{ji}(k) + \Delta W_{ji}(k) \end{cases} \tag{23}$$

$$m < 0 \text{ の時} \begin{cases} \Delta_{ji}(k) = \max(\Delta_{ji}(k-1) * \mu^{-}, \Delta_{\min}) \\ W_{ji}(k+1) = W_{ji}(k) - \Delta W_{ji}(k-1) \\ \frac{\partial \boldsymbol{E}(k)}{\partial W_{ji}(k)} = 0 \end{cases} \tag{24}$$

$$m = 0 \text{ の時} \begin{cases} \Delta W_{ji}(k) = -\mathrm{sign}(\frac{\partial \boldsymbol{E}(k)}{\partial \mathrm{W}_{ji}(k)}) * \Delta_{ji}(k) \\ W_{ji}(k+1) = W_{ji}(k) + \Delta W_{ji}(k) \end{cases} \tag{25}$$

2.4 学習曲線の振動

2.4.1 ニューラルネットワークにおける学習曲線の振動

NN 学習において，学習状況がローカルミニマムになり，学習がとまってしまい進まないとき，振幅の小さな振動をニューロンに入力することで学習が改善され，振動の効果が大きいことは, よく知られている [12]．ここで，振動に関して，従来から知られていることと，本論文で行なった事柄を，簡単に箇条書きにする．

(1) 従来からよく知られている論文は，振動を強制的に加えて（正弦波や余弦波の合成波），その効果を確かめている論文 [12] であり，学習を行なう過程で，自発的に発生した振動ではない．

(2) 学習がほとんど進行しない場合の対策について振動は有効であるとして，よく知られている．しかし，学習データと振動の関係に関しての詳しい研究は非常に少ない．学習が進行しないという場合には，学習しにくいデータが多く含まれている場合であり，学習しやすいデータが多い場合には，それほど複雑な振動は発生しない．

(3) 振動そのものの特性と学習データとの関係に関しての研究はほとんど行われていない．

(4) 任意の出力層素子への入力特性は，振幅は異なるものの，形状はどの学習データに対してもよく似ている. これは，一括更新で学習を進めているためである．

(5) 学習データをいくつかのグループに分け，多段階学習を行ない，誤差に応じた学習理知を使用することで，自発的振動は，学習が複雑になるほど活発となる．多段階学習で自発的振動を発生させ，その効果を確かめた結果, 学習しにくい関数学習が効果的に行われる．

(6) 次に，振動に関する文献は少なく，そのメカニズムは, ほとんど知られていない. 後半の研究では, 良好な学習が行なわれる時のメカニズムを教師データのグループ化を行ない検討した.

(7) 学習しにくい学習データのグループ，および，学習しやすい学習データのグループ，すなわち，誤差が常に大きいグループ，および，誤差が常に小さい学習データのグループは，それぞれ，多数派に属したり，少数派に属したりするが，グループ単位で多数派，少数派に属することで比較的ハッキリしている．しかし，誤差の少ない学習が行われる時，任意の出力層素子への入力特性は，複雑な振動特性になる．

(8) 本研究では,出力層素子への入力特性に注目し,学習データには，学習しやすいデータと学習しにくいデータがあり，さらに，複雑な振動特性になる場合には，任意の学習データに対して，学習しやすいデータと学習しにくいデータの中間的な学習データの存在が大きく影響している．詳しい検討は，4.2.4 で述べる．

2.4.2 誤差増加ベクトルおよび誤差減少ベクトル

パターン $p(=1,\cdots,P)$ に対する教師信号ベクトルを $\boldsymbol{d}^{(p)}=\{d_i^{(p)}\}(i=1,\cdots,n)$ とし，出力ベクトルを $\boldsymbol{o}^{(p)}=\{o_i^{(p)}\}(i=1,\cdots,n)$ とすると，出力誤差ベクトル $\boldsymbol{\epsilon}_{rr}^p$ は，$\boldsymbol{\epsilon}_{rr}^p=\boldsymbol{d}^{(p)}-\boldsymbol{o}^{(p)}$ となる．また，学習時のエポック t における重み係数行列を $\boldsymbol{W}(t)=\{W_{ji}(t)\}$ とする．$W_{ji}(t)$ はユニット i からユニット j への重み係数である．

ここで，誤差に応じた学習率 [1] について述べる．入力パターン p に対する学習率 η_p を以下のように設定する．

$$\eta_p=\eta\frac{(|\boldsymbol{\epsilon}_{rr}^p|)^2}{\bar{E^2}}\qquad(p=1,\cdots,P) \tag{26}$$

$$\bar{E^2}=\frac{1}{P}\sum_{p=1}^{P}(|\boldsymbol{\epsilon}_{rr}^p|)^2 \tag{27}$$

$\eta\,(>0)$ は基準の学習係数，$|\boldsymbol{\epsilon}_{rr}^p|$ は，出力誤差ベクトルの大きさ，$\bar{E^2}$ は二乗誤差の平均を表わす．多段階学習法では，あるエポックにおいて，各入力パターンに応じた学習率を用いて，そのエポックの重み更新量を計算する一括学習方式を採用する．

第 t エポックにおいて，直前のエポックよりも絶対誤差が増加，または，等しいパターンの集合を $\boldsymbol{P}^+(t)$ とする．

$$\boldsymbol{P}^+(t)\equiv\{p:|\boldsymbol{\epsilon}_{rr}^p(t)|-|\boldsymbol{\epsilon}_{rr}^p(t-1)|\geq 0\} \tag{28}$$

同様に，絶対誤差が減少するパターンの集合を $\boldsymbol{P}^-(t)$ とする．

$$\boldsymbol{P}^-(t)\equiv\{p:|\boldsymbol{\epsilon}_{rr}^p(t)|-|\boldsymbol{\epsilon}_{rr}^p(t-1)|<0\} \tag{29}$$

全パターン数は,

$$\forall t>0\ \ |\boldsymbol{P}^+(t)|+|\boldsymbol{P}^-(t)|=P \tag{30}$$

となる．ただし，式 (26)〜式 (29) の $|\cdot|$ は，誤差ベクトルの大きさを表わし，式 (30) の $|\cdot|$ は，集合の要素数を表わすものとする．本論文の計算機実験では，重み係数の更新は，すべて一括更新で行なう．したがって，第 t エポックにおける重み更新は,

$$\boldsymbol{W}(t+1)=\boldsymbol{W}(t)+\Delta\boldsymbol{W}(t) \tag{31}$$

となる．

また，$\boldsymbol{P}^{+}(t)$ に属する重み更新量の合計を $\Delta\boldsymbol{W}^{+}(t)$ とし，$\boldsymbol{P}^{-}(t)$ に属する重み更新量の合計を $\Delta\boldsymbol{W}^{-}(t)$ とすると，

$$\Delta\boldsymbol{W}^{+}(t) \equiv \sum_{p\in\boldsymbol{P}^{+}} \Delta\boldsymbol{W}^{p}(t) \tag{32}$$

$$\Delta\boldsymbol{W}^{-}(t) \equiv \sum_{p\in\boldsymbol{P}^{-}} \Delta\boldsymbol{W}^{p}(t) \tag{33}$$

となる．ここで，$\Delta\boldsymbol{W}^{+}(t)$，$\Delta\boldsymbol{W}^{-}(t)$ をそれぞれ誤差増加ベクトルおよび誤差減少ベクトル（両ベクトルの要素数は n）と呼ぶことにする．また，式 (31）の重み係数更新量とは，

$$\Delta\boldsymbol{W}(\mathrm{t}) = \Delta\boldsymbol{W}^{+}(t) + \Delta\boldsymbol{W}^{-}(t) \tag{34}$$

の関係がある．

3　多段階学習法

3.1　誤差に応じた学習係数の値の動的調整

学習時，各学習パターンに対する出力誤差は，均一に減少するのではなく，誤差が減少しやすいデータと減少しにくいデータが存在する [13]．このことに着目し，まず学習係数の基準値を設け，誤差の多いデータに対しては基準値よりも値の大きい学習係数を，誤差の少ないデータに対しては基準値よりも値の小さい学習係数を学習中に動的に設定する．これが誤差に応じた学習係数の値の動的調整である．これにより，学習回数の減少と出力層素子への入力特性が改善される [13]．ここで出力層素子への入力特性とは，学習時に任意の学習パターンに着目し，その学習パターンが各エポックで入力されたときの，任意の出力層素子への入力信号の変化である．学習係数を一定とした場合，出力層素子への入力特性は不規則な振動を示すことが多いが，誤差に応じた学習係数の値の動的調整を行なった場合には，規則的な振動になり $RMSE$ が減少する．これが出力層素子への入力特性の改善である．

入力層素子数を n 個，中間層素子数を m 個，出力層素子数は簡単のため 1 個とする．以後の議論は出力層素子が複数ある場合にも容易に拡張可能である．教師信号の集合を $\boldsymbol{D}$，個数を P，i 番目の学習パターン（入力パターン）を $\boldsymbol{x}^{(i)} = \left\{x_j^{(i)}\right\}\ (i = 1, \cdots, P : j = 1, \cdots, n)$，対応する教師信号を $d^{(i)}(i = 1, \cdots, P)$，実際の出力層素子の出力を $o^{(i)}(i = 1, \cdots, P)$ とする．入力パターン $\boldsymbol{x}^{(i)}$ に対する出力誤差 $e^{(i)}$ は，$e^{(i)} = d^{(i)} - o^{(i)}\ (i = 1, \cdots, P)$ となる．そして，目標とする教師信号 $d^{(i)}$ に対する学習係数 η_i を以下のように設定する．

$$\eta_i = \eta \frac{(e^{(i)})^2}{\bar{e^2}} \qquad (i = 1, \cdots, P) \tag{35}$$

$$\bar{e^2} = \frac{1}{P} \sum_{i=1}^{P} (e^{(i)})^2 \tag{36}$$

ここで，$\eta(>0)$ は基準の学習係数である．式 (35) および式 (36) の計算は，学習時の各エポックで実行する..

3.2　出力層素子への入力特性

任意の入力パターンベクトル $\boldsymbol{x}^{(p)}$ が入力されたときの出力層素子への入力を $O^{\mathrm{in}}(\boldsymbol{x}^{(p)})$ とする．

$$O^{\mathrm{in}}(\boldsymbol{x}^{(p)}) = \textstyle\sum_{j=1}^{m} o_j^{\mathrm{in}}(\boldsymbol{x}^{(p)}) \tag{37}$$

ここで $o_j^{\mathrm{in}}(\boldsymbol{x}^{(p)})$ は，パターン $\boldsymbol{x}^{(p)}$ が入力されたとき，出力層素子が j 番目の中間層素子から受け取る入力である．

計算機実験では，p の値はランダムに決定した後，以後は固定する．したがって添え字 p を省略し，第 t エポックにおける出力層素子への入力を $O^{\mathrm{in}}(\boldsymbol{x}(\mathrm{t}))$と表記する．

現在のエポックを t とし，T 回前までの学習においての出力層素子への入力の最大値 $O^{\mathrm{in}}_{\max}(t)$ および最小値 $O^{\mathrm{in}}_{\min}(t)$ を

$$O^{\mathrm{in}}_{\max}(t) \equiv \max_{\tau=0,\cdots,T} O^{\mathrm{in}}(\boldsymbol{x}(t-\tau)) \tag{38}$$

$$O^{\mathrm{in}}_{\min}(t) \equiv \min_{\tau=0,\cdots,T} O^{\mathrm{in}}(\boldsymbol{x}(t-\tau)) \tag{39}$$

とし，以下の $\Delta O^{\mathrm{in}}(t)$ を定義する．

$$\Delta O^{\mathrm{in}}(t) \quad \equiv \quad \frac{O^{\mathrm{in}}_{\max}(t) - O^{\mathrm{in}}_{\min}(t)}{2} \tag{40}$$

学習曲線が振動している場合，$\Delta O^{\mathrm{in}}(t)$ を第 t エポックにおける振幅とみなす．また，第 t エポックと第 $(t-1)$ エポックの間の出力層素子への入力差を $\delta O^{\mathrm{in}}(t)$ とすると，

$$\delta O^{\mathrm{in}}(t) \equiv O^{\mathrm{in}}(\boldsymbol{x}(t)) - O^{\mathrm{in}}(\boldsymbol{x}(t-1))(t = 1, 2, \cdots) \tag{41}$$

となる．T エポックの間，$O^{\mathrm{in}}(\boldsymbol{x}(t))$ が単調でなければ，

$$\mathrm{sign}(\delta O^{\mathrm{in}}(t)) \neq \mathrm{sign}(\delta O^{\mathrm{in}}(t-\tau)) \tag{42}$$

となる $\tau(= 1, \cdots, T)$ が存在する．

式 (42) が成り立つとき，第 t エポックにおける学習状態は振動状態であるとみなす．振動状態には不規則な変動も含まれる．学習曲線が収束するための良好な振動とは，規則的な振動の振幅が減衰する状態である．したがって，第 t エポックにおいて振動状態が良好であるとは，式 (43) を満たす場合である．本論文では，式 (43) の関係式を出力層素子への入力に対する振幅減少条件と呼ぶこととする．

$$\Delta O^{\mathrm{in}}(t) \quad < \quad \Delta O^{\mathrm{in}}(t-\tau) \tag{43}$$

3.3 教師データ選択方法

次に，学習時における入力データに着目し，どのような教師データに対して誤差が大きく，学習の妨げになっているのかを，様々な実験により検証した．一般に，学習における重み係数の初期値は絶対値の小さな範囲でランダムに値を設定する．すると，絶対値の大きい教師データの集合 $\boldsymbol{D}_\ell(\subseteq \boldsymbol{D})$ に対しては，重み係数の大きさが小さいが故に，出力は，シグモイド関数の中心付近しか使えず，当然，誤差は他のデータに比べて相対的に大きくなる．したがって，そのような教師パターンをまずピックアップする．また，電力需要予測や降水量推定などの時系列を扱った研究では，誤差が特に大きくなるのは急峻なピークがある部分であるといわれている [23][24]．これは，入力パターン間の距離が近いにもかかわらず，それに対応する教師データ間の距離が大きいデータに対応し，これらも選択する必要がある．

ここで注意することは，本論文では，関数近似問題に対して教師データを選択し，学習を行なっているが，一般的な時系列データを学習するニューラルネットワークや実数値を含むパターン学習に対しても応用可能である．

それら教師データの集合を $\boldsymbol{D}_d(\subseteq \boldsymbol{D})$ とする．任意の 2 つの入力パターン $\boldsymbol{x}^{(p)}$ と $\boldsymbol{x}^{(q)}$ について，

$$g(\boldsymbol{x}^{(p)}, \boldsymbol{x}^{(q)}) \equiv \frac{|\boldsymbol{d}^{(p)} - \boldsymbol{d}^{(q)}|}{\| \boldsymbol{x}^{(p)} - \boldsymbol{x}^{(q)} \|} \tag{44}$$

の値が大きい場合に，$\boldsymbol{d}^{(p)}$ と $\boldsymbol{d}^{(q)}$ を集合 $\boldsymbol{D}_d(\subset \boldsymbol{D})$ の要素とする．$\|\cdot\|$ はユークリッドノルムを表わす．

本論文では，すべての教師データを最初から用いるのではなく，誤差が大きいと考えられるデータ $(\boldsymbol{D}_\ell \cup \boldsymbol{D}_d)$ の選択を最初に前処理として行なう．教師データを段階的に追加することにより，計算時間の短縮が期待できる．

3.4 効率的多段階学習法

誤差に応じた学習係数の値の動的調整，出力層素子への入力特性，教師データの選択について ⟨3.3⟩ までに述べた．それらに基づく効率的多段階学習法を図 4 に示し，以下に，効率的多段階学習について述べる．学習結果が重みの初期値に依存しにくく，総合的な学習時間を短縮するために最も重要となる，誤差に応じた学習係数の値の動的調整，振動特性の利用，教師データの選択をしたうえでの多段階学習について述べる．ここで総合的な学習時間とは，パラメータの値の調整や学習のやり直しも含め，ある対象を学習するために要する総計算時間のことである．

本論文では，入出力パターンをベクトルとして扱える問題を対象とし，ベクトルの大きさやベクトル間の距離を学習に利用する．したがって，それらを利用できない問題，例えば 0 および 1 の組み合わせが出力パターンとなる識別問題などには適用できないが，関数近似問題や時系列パターンを用いた予測 · 推定問題など，広い範囲で適用可能である．

1. 教師データを s 段階に分割する．第 q 段階で用いる教師データは $\boldsymbol{D}_\ell^q \cup \boldsymbol{D}_d^q$ であり，以下の式 (45) を満たすように各段階で用いるデータを決定する．

$$\begin{aligned} (\boldsymbol{D}_\ell^{q-1} \cup \boldsymbol{D}_d^{q-1}) \subset (\boldsymbol{D}_\ell^q \cup \boldsymbol{D}_d^q)(q = 2, \cdots, s) \\ \boldsymbol{D}_\ell^s \cup \boldsymbol{D}_d^s = \boldsymbol{D} \end{aligned} \tag{45}$$

2. s 段階に分けた教師データを，追加しながら学習を行なう．すなわち，第 1 段階で，全教師データの約 $\frac{1}{s}$，第 2 段階で全教師データの約 $\frac{2}{s}$，以下同様とし，学習回数は各段階で同数とする．

3. 学習中は，常に誤差に応じた学習係数の値の動的調整を行ない，重み係数の値は次の段階に引き継ぐ．

4. 最終の第 s 段階ではすべての教師データを用いて学習する．

5. 第 s 段階開始時出力層素子への入力特性が振幅減少条件（式 (43)）を満足していない場合には，学習が失敗する傾向が強いので，失敗と判断し，新たな初期値を設定し，学習を開始する．この場合，やり直しではあるが，学習の早い段階で判定できるので，無駄な計算時間（第 3 段階目の初期で学習を終了する場合，学習時間は全学習時間の $\frac{1}{2}$）を極力抑えることができる．さらに，良好な振動で，振幅が減少する場合，すなわち，最終段階学習時において，振幅減少条件が成立する場合には，学習が成功する確率が高いので，重み係数を引き継ぎそのまま学習を継続する．学習回数は各段階で同数とするため，学習時間の合計は，従来の方法と比較すると，$\frac{s+1}{2s}$ （実験で用いた $s = 3$ では $\frac{2}{3}$）となる．

6. 第 s 段階学習開始時において，振幅減少条件が成立しなくとも，次のことが考えられる．シグモイド関数を sig とする．出力層素子への目標とする入力は，$\mathrm{sig}^{-1}(\boldsymbol{d}^{(p)})$ となる．最初に決定した p 番目のパターンについて，目標出力を達成するためには，出力層素子への入力が $\mathrm{sig}^{-1}(\boldsymbol{d}^{(p)})$ に収束する必要がある．振動状態にあり，さらに振動の範囲内に出力層素子への目標とする入力がある時，すなわち，

$$O^{\mathrm{in}}_{\mathrm{min}}(t) < \mathrm{sig}^{-1}(\boldsymbol{d}^{(p)}) < O^{\mathrm{in}}_{\mathrm{max}}(t) \tag{46}$$

   という条件が成り立つ時，振幅減少条件が成立しなくても，良好な学習が可能である．また，振動の範囲内に出力層素子への目標とする入力がなくとも，$\mathrm{sig}^{-1}(\boldsymbol{o}^{(p)})$ が $\mathrm{sig}^{-1}(\boldsymbol{d}^{(p)})$ に近い (近いとは，出力層素子への入力特性において，目標値に最も近い入力と目標値との差が振動の振幅より小さい場合を言う) ほど誤差は少なくなる．ここで，式 (46) を満たす場合，または，振動範囲が目標値に近い場合を，目標値捕捉条件を満たすと表現することにする．振幅減少条件を満たさなくても，目標値捕捉条件を満たす場合は，良好な学習が期待できるため，学習を継続する．

7. 振幅減少条件 (式 (43)) を満足するかまたは振幅減少条件を満たさなくとも目標値捕捉条件を満足している場合には，設定された $RMSE$ の値によって，学習を終了するか学習を継続するかを決定する．すなわち，設定された $RMSE$ の値よりも小さい場合には，学習を終了し，そうでない場合には，重み係数を引継ぎ最終段階の学習を継続する．

ここで，効率的学習法の要点は，最後に述べた．出力層素子への入力特性が，振動を繰り返し，さらにその振幅が減少している場合は良好な学習が行われる．また，振幅が減少していなくとも，目標値に近い振動をしている場合（目標値を振動範囲内に含んでいる）には学習が成功するということを意味している．この例は，関数近似問題を例とした計算機実験１で，確かめられる．

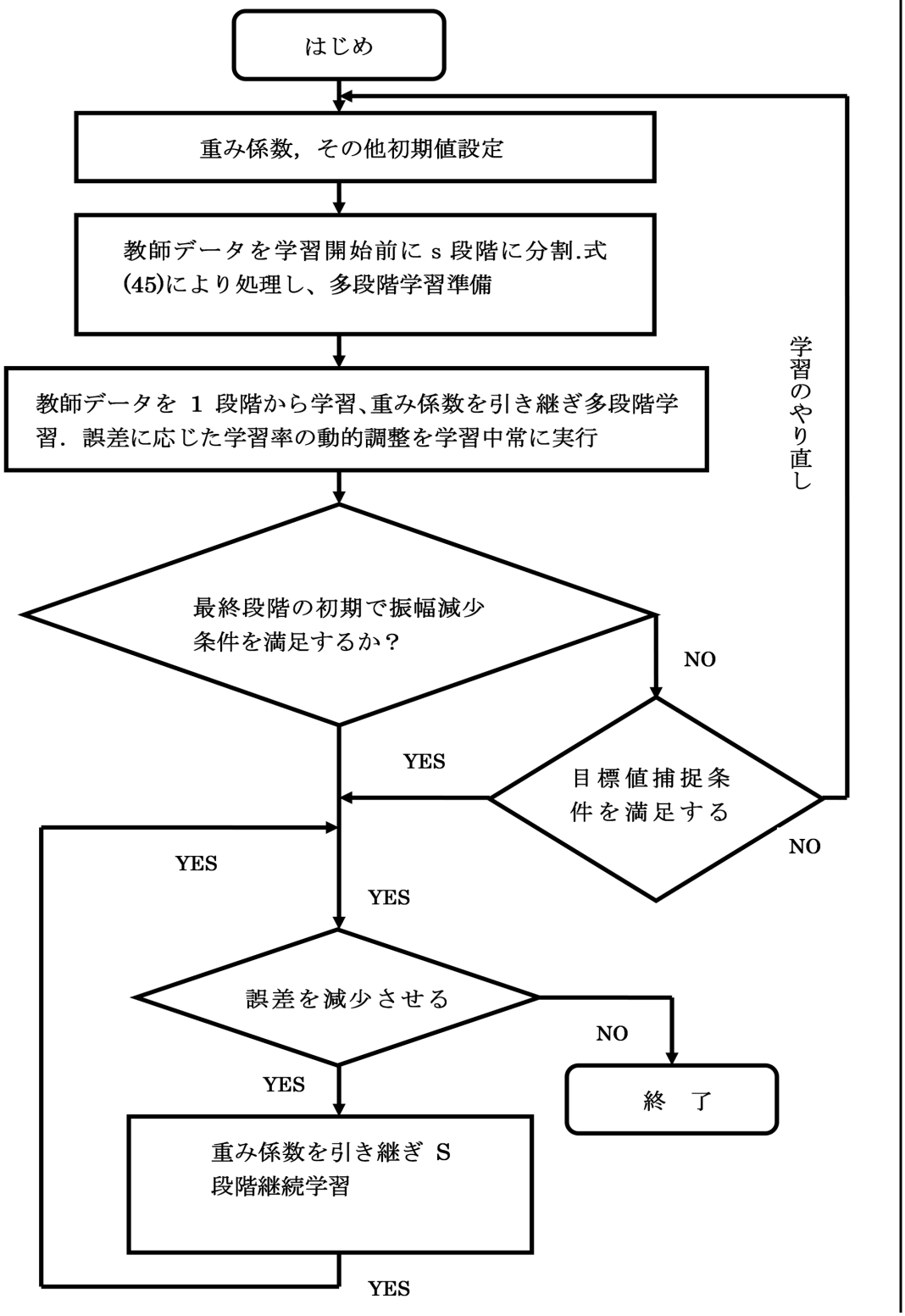

図 4: The method of general learning

3.5 多段階 QPROP 法

本論文では，教師データの選択および誤差に応じた学習係数の使用による多段階学習によって，学習時間，学習精度を向上させることが期待できる [25]．ここでは，多段階学習法の学習則に QPROP 法を組み込んだ多段階 QPROP 法について述べる．

- 学習データには学習が困難なデータと容易なデータに対し，1 番困難なデータを 1 段階とし s 段階まで学習データを分割する．2 段階の学習では，1 段階で学習されたデータも含めて学習を行なう．３段階以降も同様である．
- 学習データは，困難な学習データから学習を開始する．重みの更新は，一括更新ですべてのエポックで行なう．
- 第 k エポックにおいて，1 段階のすべての学習データに対して，誤差に応じた学習係数を用い重み更新量を加算する．
- ここで，QPROP 法に従って，第 $k-1$ エポックで保存されている重み更新量を用い，式 (17) および式 (18) から学習係数を決定する．
- 慣性係数 $\alpha_{ji}^{(k)}$ に対しては，式 (19) を計算する．式 (20) および式 (21) に応じて $\alpha_{ji}^{(k)}$ を決定する．ただし，最大慣性係数 $\mu(0.95)$ を設定し，大きすぎないようにしている．さらに，重み変更抑制係数 $\lambda(0.005$ と小さい) を用いている．
- ここで，１エポックの重み更新量が決定される．
- 重み係数の更新を実行する．
- 第 k エポックにおいて，すべて重みの更新量の決定後，第 $k+1$ エポックの重み更新量に対する学習係数と慣性係数を決定するために，第 k エポックの重みの一括更新量を記憶する．
- 第 $k+1$ エポックの学習を開始し，順次同様な計算を繰り返し，1 段階の学習を終了する．
- 重みを引き継ぎ，2 段階について同様の計算を行なう．注意することは，2 段階においては，1 段階で使用した学習データを含んで学習を行なうことである．
- s 段階終了まで，以上で述べた方法で学習を継続する．

3.6 多段階 RPROP 法

多段階 RPROP 法は，多段階 QPROP 法と組み込む前は全く同様で，教師データの選択および誤差に応じた学習係数の使用による多段階学習である．ここでは，多段階学習法の学習則に RPROP 法を組み込んだ多段階 RPROP 法について述べる．QPROP 法と異なる点は，慣性項に対する処理は行っていないことである．

- 学習データには学習が困難なデータと容易なデータに対し，1 番困難なデータを 1 段階とし s 段階まで学習データを分割する．2 段階の学習では，1 段階で学習されたデータも含めて学習を行なう．3 段階以降も同様である．
- 学習データは，困難な学習データから学習を開始する．重みの更新は，一括更新ですべてのエポックで行なう．
- 第 k エポックに対し，1 段階のすべての学習データに対して，誤差に応じた学習係数を用い重み更新量を加算する．
- RPROP 法は，基本的には，振動を抑えるために前回の重み更新量を記憶しておくことと，前回の重み更新量の符号と今回の重み更新係数の符号を考慮し，5 個のパラメーターを条件に応じて調整することにより，振動を抑えた学習を行なう方法である．
- 式 (22) に基づき，$m = \frac{\partial \boldsymbol{E}(k-1)}{\partial W_{ji}(k-1)} * \frac{\partial \boldsymbol{E}(k)}{\partial W_{ji}(k)}$ を計算する．
- m の符号（$m > 0$，$m < 0$，$m = 0$）によって重み更新量が異なる．また，更新量が大きすぎないように $\Delta_{\max}$，小さすぎないように $\Delta_{\min}$ を設定している．すなわち，$m > 0$ の場合には，更新量を多めにとって計算をさらに進め，$m < 0$ の場合には，重み係数を前の状態に戻す操作をしている．$m = 0$ の場合には，重みの更新値は 0 である．また，重み係数更新量とそれに対応する $\Delta_{ji}(k)$ および $\Delta W_{ji}(k)$ は，第 $k+1$ エポックの計算のためにつねに一括更新量を記憶する．
- 第 $k+1$ エポックの学習を開始し，順次同様な計算を繰り返し，1 段階の学習を終了する．
- 重みを引き継ぎ，2 段階について同様の計算を行なう．注意することは，2 段階においては，1 段階で使用した学習データを含み学習を行なうことである．
- s 段階終了まで，以上で述べた方法で学習を継続する．

4　関数近似問題を例とした計算機実験

4.1　関数近似問題を例とした計算機実験１

本章で提案している多段階学習法の有効性を示すために，ここでは関数近似問題を例として計算機実験を行なう．関数近似問題を選択した理由は，学習する際の難易度を関数の選択によって設定でき，未学習データに関して実験の検証がしやすいためである．また，いくつかの関数はベンチマーク的存在であり，他手法との比較が可能である．さらに，教師データ集合 $\boldsymbol{D}_d$ (⟨3.3⟩) を式 (44) に基づいて決定する代わりに，偏微分値を利用できるとともに，関数近似問題自体の応用が広い．

学習が困難な関数の例として，式 (47) が知られている．この関数の学習を行なうと，学習曲線が収束しない，あるいは，収束したとしても多大な計算時間を要する [13][15][17]．その近似対象関数を図 5 に示す．

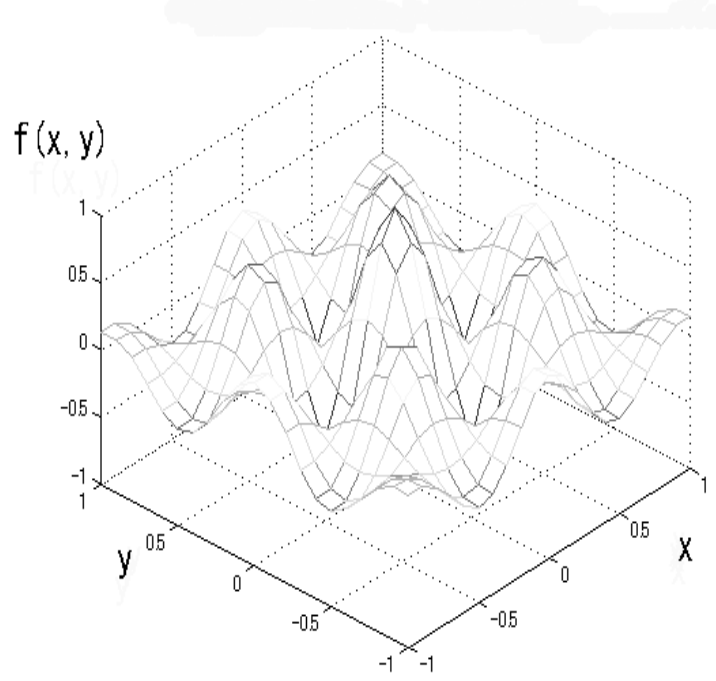

図 5: $f(x,y) = \exp(-x^2 - y^2)\cos(2\pi x)\cos(2\pi y)$.

$$f(x,y) = \exp(-x^2 - y^2)\cos(2\pi x)\cos(2\pi y) \tag{47}$$
$$(-1.0 \leq x \leq 1.0, -1.0 \leq y \leq 1.0)$$
$$(-1.0 \leq f(x,y) \leq 1.0)$$

さらに，シグモイド関数を用いた従来の NN で近似可能な式 (48) に示す関数を近似対象とし，従来法と提案手法による学習時間の比較実験を行なう．

$$f(x,y) = \frac{\sin(\pi x)\cos(\pi y)+1}{2} \tag{48}$$
$$(-1.0 \le x \le 1.0, -1.0 \le y \le 1.0)$$
$$(0.0 \le f(x,y) \le 1.0)$$

4.1.1 教師データ

式 (47) では，学習すべき領域は $-1.0 \le x, y \le 1.0$ として示されている．しかしながら，学習上非常に重要な教師データ $\boldsymbol{D}_d$ に相当するデータのほとんどは，領域 $-0.5 \le x, y \le 0.5$ に含まれるため，教師データは $-0.5 \le x, y \le 0.5$ の範囲から作成した．さらに，教師データは，x,y 方向の刻み幅を 0.1 として領域を 11×11 の格子に分割し，それらの格子点 121 点に対する値を教師データの集合 $\boldsymbol{D}$ とし，文献 (15) と等しくする．また，未学習のテストデータについては，全領域を 20×20 の格子に分割し，その格子点 400 点を用いる．式 (47) を図 6 に示す．図 6 において，隣り合う教師データは，実線で結んで表示している．

式 (48) では，学習すべき領域は $-1.0 \le x, y \le 1.0$ である．教師データは，x,y 方向の刻み幅を 0.2858 として領域を 7×7 の格子 (49 格子点) に分割 し，教師データの集合 $\boldsymbol{D}$ とした．式 (48) の概観を図 7 に表示する．

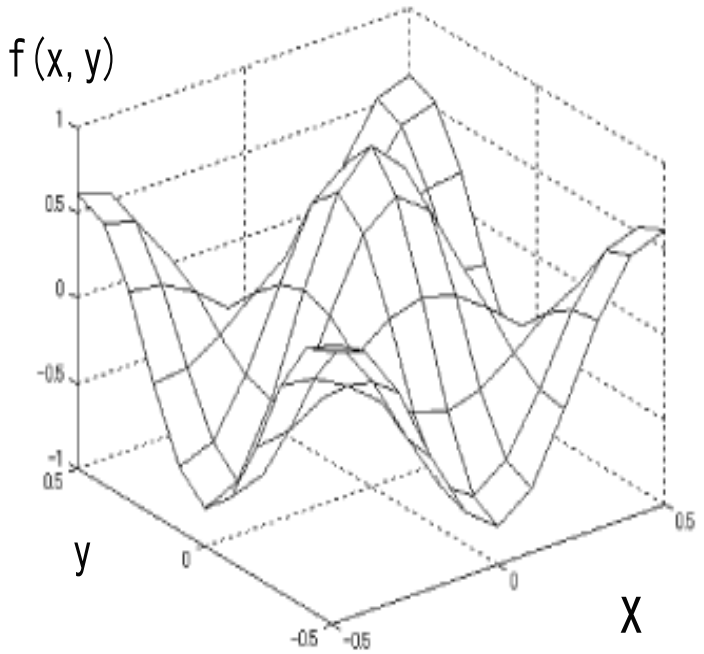

図 6: $f(x,y) = \exp(-x^2 - y^2)\cos(2\pi x)\cos(2\pi y)$.

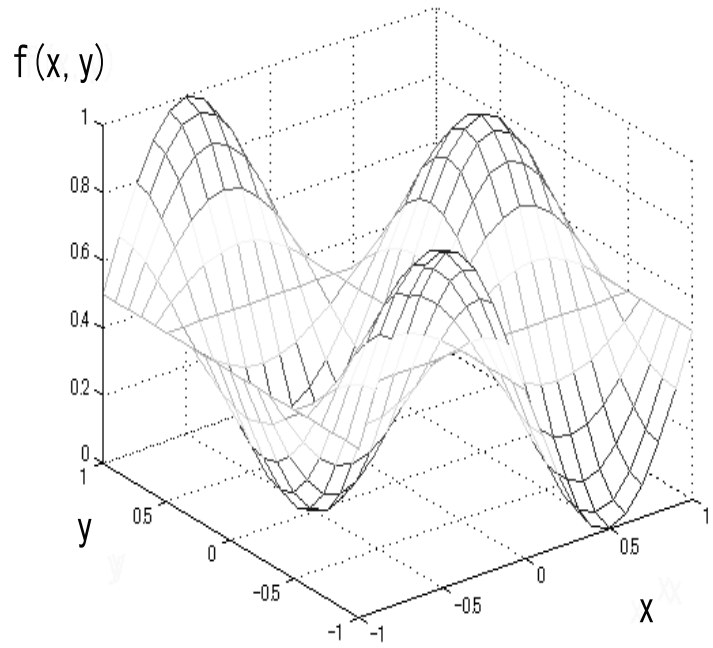

図 7: $f(x,y) = \frac{\sin(\pi x)\cos(\pi y)+1}{2}$.

4.1.2 教師データの選択

関数 (47) を学習させるための教師データの選択方法について述べる．式 (44) によると隣りあうデータであるにもかかわらず，$f(x,y)$ の値が大きく異なるデータを教師データとして選択することになるが，ここでは関数形が既知であるので x，y 両方向の偏微分値 $f_x(x,y)$ または $f_y(x,y)$ で代用できる．関数 (47) は，原点で最大値を示し，$f_x(x,y)=0$，$f_y(x,y)=0$ である．原点から離れるにしたがって，絶対値の大きさは，振動しながら減少する．

ここでは教師データを 3 段階に分け (s=3)，多段階学習を行なう．$\boldsymbol{D}^1_\ell \cup \boldsymbol{D}^1_d$ は，全体の約 30% の教師データを用いる．$\boldsymbol{D}^1_\ell$ および $\boldsymbol{D}^1_d$ （$f_x(x,y)$ または $f_y(x,y)$ の値を利用して選択）は，関数の概観を把握するいう意味で重要である．$\boldsymbol{D}^1_\ell \cup \boldsymbol{D}^1_d$ に含まれるデータ数は，最終的に 29.6%（37 個）になった．これらのデータには，$|f(x,y)| \geq 0.68$ を満たす教師データ 13 個，$|f_x(x,y)| \geq 4.3$ を満たす教師データ 22 個，$|f_y(x,y)| \geq 5.6$ を満たす教師データ 4 個が含まれ，そのうち 2 個が重複しているため，全体で 37 個となる．

第 2 段階では，絶対値の大きい教師データ数は増やさず，教師データ選択を行なった．絶対値の大きい教師データは，第 1 段階ですでに多く含まれているために，第 2 段階では，偏微分値の絶対値の大きな教師データのみ（$|f_x(x,y)| \geq 2.0$）を中心に選択した．すなわち，これらのデータには，$|f(x,y)| \geq 0.68$ を満たす教師データ 13 個（$\boldsymbol{D}^2_\ell = \boldsymbol{D}^1_\ell$），$|f_x(x,y)| \geq 2.0$ を満たす教師データ 58 個，$|f_y(x,y)| \geq 5.6$ を満たす教師データ 4 個が含まれ，そのうち 6 個が重複しているため，全体で 69 個となる．

第 3 段階における $\boldsymbol{D}^3_d$ は，全教師データを使用する．ここで，第 1 段階では，全データの 29.7%，第 2 段階では，全データの 57%，最後の第 3 段階で 100% の教師データが用いら

れることになる.

関数 (48) を学習させるための教師データの選択方法については，表 1 にまとめる.

表 1: The Method of Learning

Initial weight vector is modified 5 times. One output layer unit. Updating a weight vector should be done every 1 epoch.	
Learning data	Learning domains are $-1.0 \leq x \leq 1.0$, $-1.0 \leq y \leq 1.0$. Learning data is conditioned equally as the reference 13. All the domain is divided into 7×7 grid. All the learning data($\boldsymbol{D}$)is divided into 3 parts before learning. $0 < D^i < 1, D^i \in \boldsymbol{D} \quad (D^i = f(x_i, y_i); i = 1, \cdots, 49)$
1st step	$\boldsymbol{D}_\ell^1 = \{f(x_i, y_i) \mid \lvert f(x_i, y_i)\rvert \geq 0.7\}$（11 learning data） $\boldsymbol{D}_d^1 = \{f(x_i, y_i) \mid \lvert f_x(x_i, y_i)\rvert \geq 1.4$ or $\lvert f_y(x_i, y_i)\rvert \geq 1.5\ \}$(6 learning data) $\lvert \boldsymbol{D}_\ell^1 \cup \boldsymbol{D}_d^1 \rvert$ =1（1 overlapped） 16 learning data are selected.（about 33%）
2nd step	$\boldsymbol{D}_\ell^2 = \{f(x_i, y_i) \mid \lvert f(x_i, y_i)\rvert \geq 0.7\}$（11 learning data） $\boldsymbol{D}_d^2 = \{f(x_i, y_i) \mid \lvert f_x(x_i, y_i)\rvert \geq 1.2$ or $\lvert f_y(x_i, y_i)\rvert \geq 1.0\ \}$（21 learning data） $\lvert \boldsymbol{D}_\ell^2 \cup \boldsymbol{D}_d^2 \rvert$ =2（2 overlapped） 30 learning data are selected. （about 61%）
3rd step	49 learning data are selected.(100%)
Each step	Every step are learned for 20,000 times.
Conventional method	Learn 60,000 times with all the learning data and learning coefficient 1.0. Initial weight vector is modified 5 times.

4.1.3 比較対象手法と実験の諸設定

計算機実験では提案手法の有効性を検証するために，従来手法との比較を行なう．関数 (47) に対する提案手法および従来法を用いた場合の NN の構成は，シグモイド素子を用い，入力層素子数 2 個，中間層素子数 9 個，出力層素子数 1 個からなるフィードフォワード型の 3 層構造のネットワークとする．中間層素子数は予備実験より決定した．さらに，重み係数の更新は，両手法とも 1 エポック毎の一括更新方式を用いた．提案手法の学習回数は，1 段階目 2333 エポック，2 段階目 2333 エポック，3 段階目 2334 エポックの合計 7000 エポックとし，従来法は 7000 エポックの学習で全教師データを常に用いる．実験に用いる計算機は，OS:Windows XP, CPU:Pentium 4, 3.0GHz, RAM:2GB である．また，学習係数は，提案手法における基準の学習係数を $\eta = 0.8$ とし（式 (35) 参照），従来法は学習係数 0.8 で一定とした．学習結果を評価する $RMSE$ に関しては，教師データに対する値を用い，重み係数の初期値を 5 通り設定した場合の平均値を用いる．重み係数の初期値は $[-0.01, 0.01]$ の範囲でランダムに設定する．

関数 (48) に対する提案手法および従来法を用いた場合の NN の構成は，入力層素子数 2 個，中間層素子数 10 個，出力層素子数 1 個からなる各素子にシグモイド素子を用いたフィードフォワード型の 3 層構造のネットワークとする．重み係数の更新は，両手法とも 1 エポック毎の一括更新方式を用いる．提案手法のエポック数は，1 段階目，2 段階目，3 段階目ともに 20000 エポックとする（従来法は 60000 エポック）．学習係数の値を 1.0（提案手法の基準の学習係数 $\eta = 1.0$，従来の学習係数 1.0 で一定）とする以外は，$RMSE$ の取り扱いと評価方法，重み係数の初期値設定方法を，関数 (47) を扱う場合と同じとする．

4.1.4 計算機実験結果 1

関数 (47) に対する提案手法および従来法を用いた場合の学習データおよび未学習データに関する結果を表 2，表 3 に示す．関数 (47) に関する未学習データについては，文献 [15] を参考とし，x, y 両領域 $(-0.5 \leq x \leq 0.5, -0.5 \leq y \leq 0.5)$ に対し,10×10 の格子に分割し (学習データを除く)，その中の格子点 100 点を用いた．提案手法を用いた場合に，学習データおよび未学習データに対する $RMSE$ が最も小さくなることがわかる．表 2 には提案手法に慣性項を付け加えた場合の結果を含み，表中の数値は重みの初期値をランダムに設定した５回分の実験結果である．

提案手法に対しては，慣性項を付け加えた場合を含め，学習のやり直しは１度も生じなかった．従来法では学習曲線がほとんど収束しないため，実験結果５例を得るために数 10 回の重みの初期値設定を必要とした．慣性係数の値を変えても同じ状況であった．数 10 回の実験の中には，学習回数を 7000 エポックから 15000 エポック，30000 エポックに増やした場合を含む．表 2 には学習終了時点（規定のエポック終了時点）で $RMSE$ 値が小さかった 5 例を示している．

提案手法，従来法 $(\alpha = 0.0)$，提案手法に慣性項を設けた場合 $(\alpha = 0.5)$ に対し，学習の第 3 段階における出力層素子への入力特性を，それぞれ，図 8〜図 10 に示す．これらは表 2 に用いた 5 例のうちの最初の実験例に関する図であり，図 8，図 10 は第 3 段階の 2334 エポック分，図 9 はそのうちの一部（900 エポックから 1100 エポックまでの 200 エポック分）を示している．図 9 の 200 エポック分は，第 3 段階中，振幅減少条件を満足していない部

表 2: RMSE for Proposed Methods and Traditional Methods (learning data of the function(47))

Proposed method	$RMSE$ Mean	Maximum value	Minimum value
	0.052	0.089	0.033
Learning time	Mean	7 minutes 40 seconds	

Proposed method by using α Inertia coefficient α	$RMSE$ Mean	Maximum value	Minimum value
$\alpha = 0.5$	0.089	0.129	0.058
$\alpha = 0.8$	0.110	0.186	0.083
Learning time	Mean	7 minutes 40 seconds	

Conventional method	$RMSE$ Mean	Maximum value	Minimum value
$\alpha = 0.0$	0.146	0.176	0.124
$\alpha = 0.8$	0.382	0.574	0.163
Learning time	Mean	12 minutes 40 seconds	

分である．図 8 における実線は，出力層素子への入力の目標値 1.01（誤差が多く残るデータを乱数で選択）であり，図 9 および図 10 には示していないが同じ目標値を用いている．ここで，目標値は，学習が良好に行われ，誤差がないときの出力層素子への入力値である．目標値と出力層素子への入力特性によって，任意の教師データに対する学習の良好さを知ることができる．

関数 (48) に対する提案手法および従来法を用いた場合の学習データおよび未学習データに関する結果を，それぞれ，表 4，表 5 に示す．

学習のやり直しはなく，それぞれ 5 回の実験結果から得られた数値である．提案手法に対する，1 回目の実験の第 3 段階における出力層素子への入力特性を図 11 に示す．図 11 における出力層素子への入力の目標値は，0.000 である．

4.1.5 実験結果 1 の考察

計算機実験に基づき，学習の性能と計算時間について考察する．

学習の性能

表 2 より，従来法では $RMSE$ が 0.120 より小さくなることはなく，提案手法の最大値 (0.089) と比較してもかなりの差がある．表 2 の結果を得るために数十回の実験が必要であったことから，実質的には従来法では関数 (47) の学習は不可能であるといえる．また、学習

表 3: RMSE for Proposed Methods and Traditional Methods (test data of the function(47))

Proposed method	$RMSE$ Mean	Maximum value	Minimum value
	0.099	0.147	0.069
Learning time	Mean	7 minutes 40 seconds	

Proposed method by using α Inertia coefficient α	$RMSE$ Mean	Maximum value	Minimum value
$\alpha = 0.5$	0.199	0.263	0.145
$\alpha = 0.8$	0.267	0.446	0.132
Learning time	Mean	7 minutes 40 seconds	

Conventional method	$RMSE$ Mean	Maximum value	Minimum value
$\alpha = 0.0$	0.311	0.376	0.281
$\alpha = 0.8$	0.838	0.348	1.273
Learning time	Mean	12 minutes 40 seconds	

データおよび未学習データともに提案手法、提案手法に慣性係数を使用した場合、従来法の順に RMSE 値が大きくなる結果となり、提案手法の学習誤差が小さいことがわかる．

提案手法は，関数 (47) の学習が可能であり，図 8 の提案手法による特性では，振動は減衰し，ほとんどすべてのエポックで目標値を振動範囲内に含んでいるのに対し，図 9 の従来法では目標値（1.01）から離れた範囲で振動し，振幅は増減を繰り返している．提案手法と従来法との相違は，教師データの選択の仕方と誤差に応じた学習係数の動的調整の 2 点であり，これらの 2 点の相違が特性の相違として生じたといえる．

関数 (48) は，従来法で学習可能であり（表 4 参照），提案手法による学習結果である図 11 では，振幅減少条件を満たしているので，第 3 段階の学習に進んでいる．それに対し，図 8 の学習（関数 (47)）では，第 3 段階開始時において，振幅減少条件を満足していない．つまり，図 8 は第 3 段階の学習であり，この段階に進むかどうかは図 8 の y 軸付近の様子で判定される．y 軸付近では増減していて振幅減少条件は満たされず，目標値は 1.01 であるから目標値捕捉条件は満たしている．したがって，振幅減少条件だけでは学習はやりなおしになっていたはずである．この例は学習が成功した一例である．目標値捕捉条件を設けたことにより，学習は継続され，良好な最終結果を得ている．このことは，目標値捕捉条件を設けたことは有効であることを示している．

つぎに慣性項の効果について検討する．図 10 では，目標値捕捉条件を満たしているので第 3 段階の学習に進んでいる．しかしながら，振動現象の振幅が大きく，かつ，不規則であり，このことが誤差を大きくしていると考えられる．したがって，式 (47) のような複雑な形状の関数を学習させる場合には，慣性項は有用ではない．

表 4: Learning Time and RMSE (learning data of the function(48))

Proposed Method	$RMSE$ Mean	Maximum value	Minimum value
	0.046	0.069	0.028
Learning time	Mean	24minutes 35 seconds	

Conventional method Inertia coefficient	$RMSE$ Mean	Maximum value	Minimum value
$\alpha = 0.0$	0.067	0.117	0.041
$\alpha = 0.5$	0.059	0.093	0.035
$\alpha = 0.8$	0.050	0.093	0.034
Learning time	Mean	39 minutes 56 seconds	

学習時間

関数 (48) の提案手法では，学習時間は，24 分 35 秒となり，従来法に比較するとほぼ 62% という良好な結果となり，(表 4 参照) 計算時間は，ほぼ推定値通りとなった．⟨3.4⟩ で示した推定値 67% よりわずかに少ない時間で学習が行なわれた．このことは，提案手法における教師データの選択が有効で，学習時間が改善されたことを示す．

関数 (47) についても同様な結果となり，表 2 に示すように，約 60% の計算時間となっている．

ここで，提案手法を用いた場合，学習のやり直しを含めた総合的な学習時間について，本手法でも 3 回の学習時間のやり直しがあり，1 回成功した場合を想定し，表 2 の結果を基として計算時間の比較をする．問題を簡単化するために，従来法で 1 回学習する時間を 1 と仮定し，計算を行なう．本手法の場合，$\frac{2}{3}$，それに加えて，やり直し 3 回分はすべて第 2 段階で終了しているので，$(\frac{1}{9} + \frac{2}{9})$ の 3 倍で，合計 1 となる．したがって，総計算時間は，$\frac{5}{3}$ である．一方，従来法では，関数 (47) に対して 5 回の結果を得るのに数十回の学習が必要であった．これを大ざっぱに 5 回に対して 50 回の学習が必要であるであると仮定すると，1 つの結果を得るのに平均 10 回分の学習，すなわち，従来法による総計算時間は 10 となる．したがって，総計算時間は，$\frac{\frac{5}{3}}{10} = \frac{1}{6}$ に短縮できる．ここで，表 2 のデータに基づくと，1 つの学習結果を得るために，従来法では，127 分程度 (12 分 40 秒の 10 倍) に対し，提案手法では 21 分程度 (12 分 40 秒の $\frac{5}{3}$) ですむことになる．

本章では，学習データには学習が困難なデータと容易なデータがあり，学習困難なデータから多段階的に学習を行なうこと，しかも，誤差に応じた学習率を使用することによって学習時間の短縮，高い学習精度を実現した．しかし，BP 法に他の学習法を組み込むこと

表 5: Learning Time and RMSE (test data of the function(48))

Learning domain are $-1.0 \leq x \leq 1.0,\ -1.0 \leq y \leq 1.0$. All the domain is divided into 20×20 grid. 400 test data are selected.			
Proposed Method	$RMSE$ Mean	Maximum value	Minimum value
	0.067	0.078	0.060
Learning time	Mean	24minutes 35 seconds	
Conventional method Inertia coefficient	$RMSE$ Mean	Maximum value	Minimum value
$\alpha = 0.0$	0.080	0.126	0.054
$\alpha = 0.5$	0.077	0.103	0.056
$\alpha = 0.8$	0.076	0.118	0.058
Learning time	Mean	39 minutes 56 seconds	

による性能評価や振動現象のメカニズムは問題として残る．次章では，組み込み効果や重み更新ベクトルを導入することにより，振動メカニズムについての報告をする．

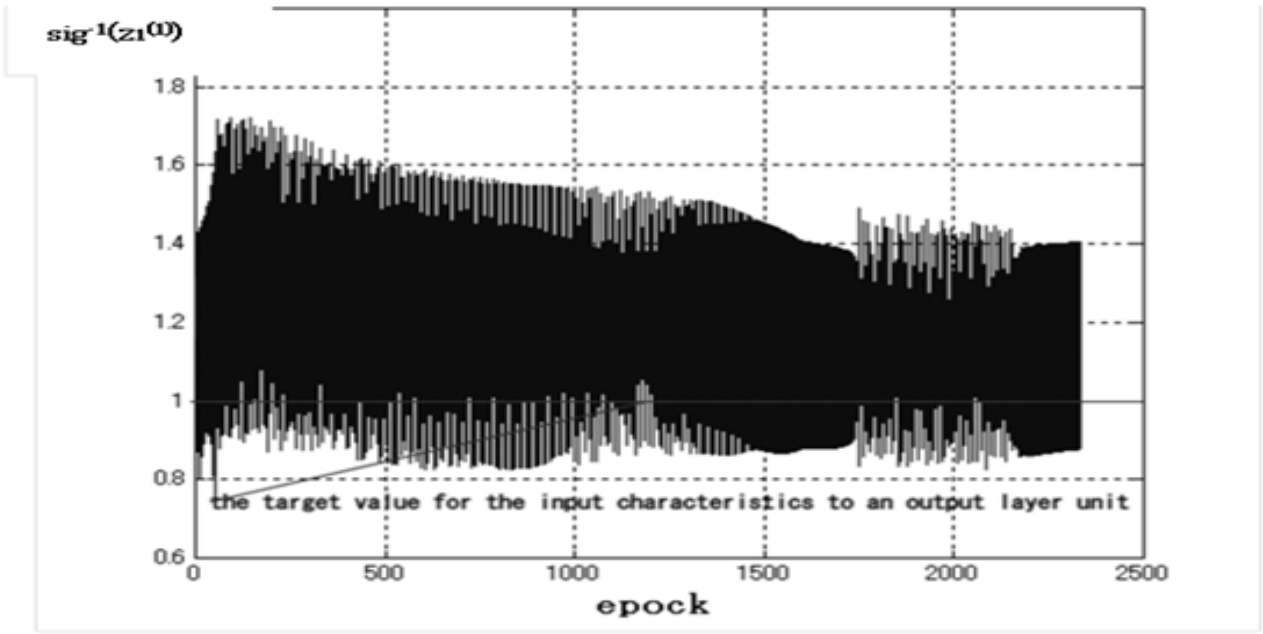

図 8: Input characteristics for the proposal methods in the output layer.

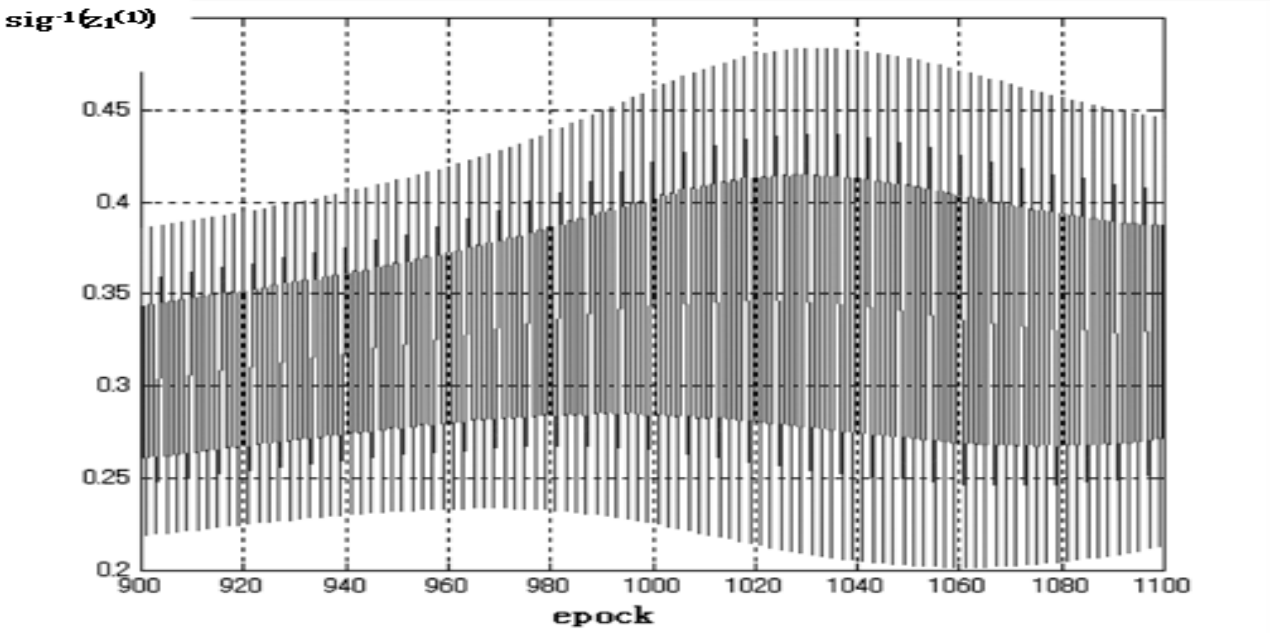

図 9: Input characteristics for general method in the output layer ($\alpha = 0$).

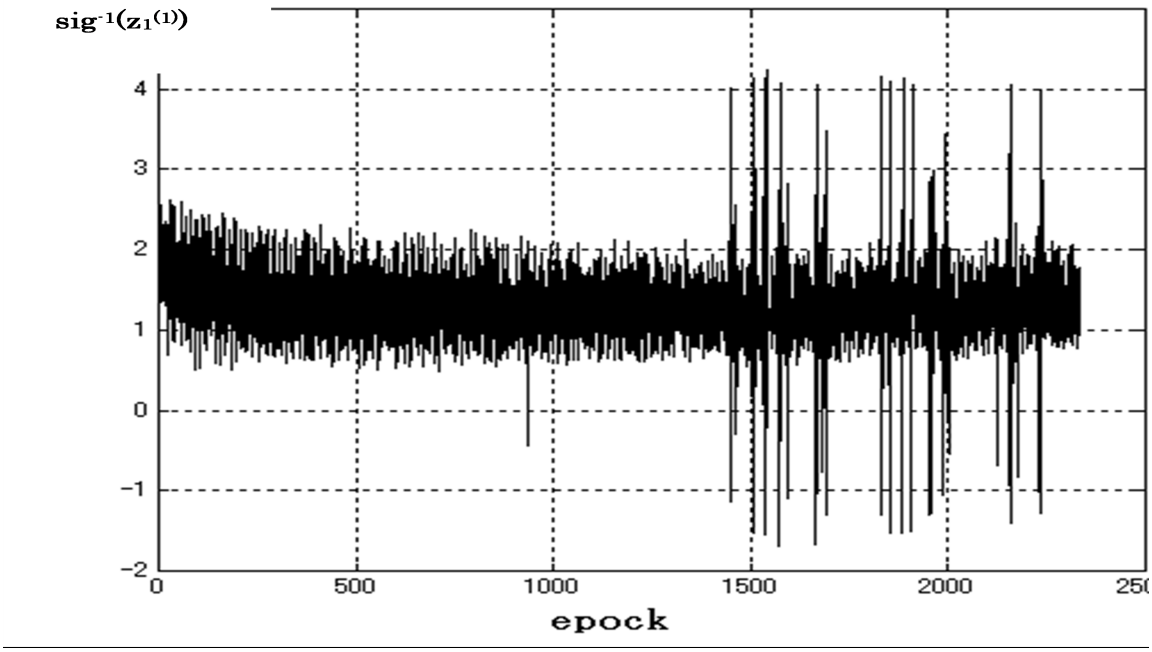

図 10: Input characteristics for $\alpha = 0.5$ in the output layer.

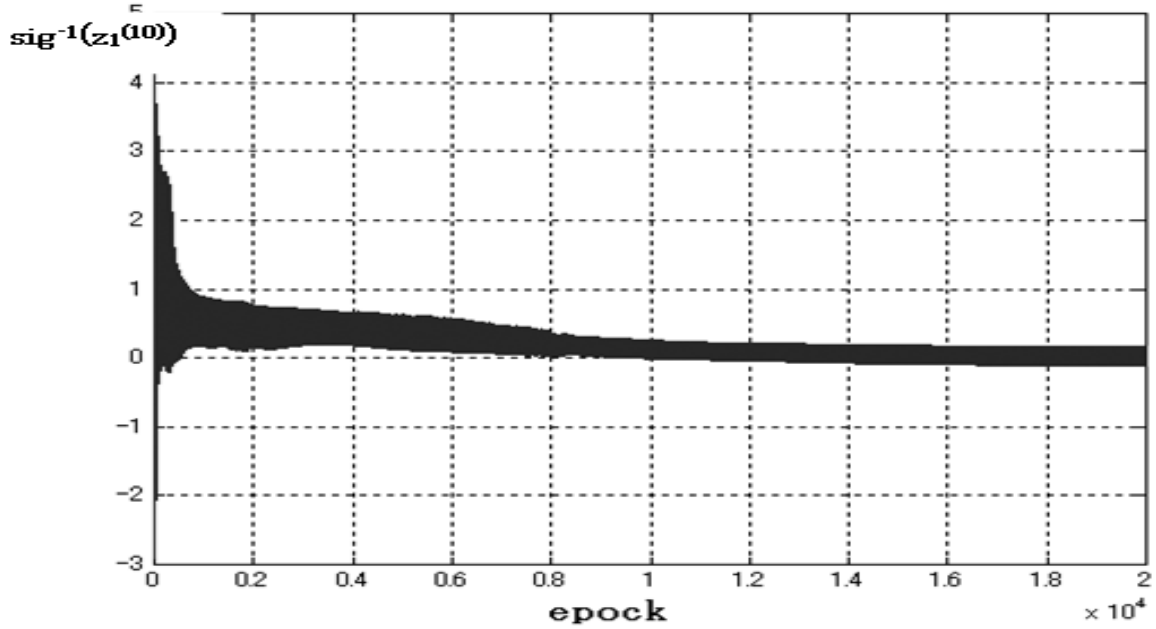

図 11: Input characteristics for the proposal methods in the output layer.

4.2 関数近似問題を例とした計算機実験２

ここでは，関数最適化問題 [42] のテスト関数としてよく用いられるいくつかの関数を関数近似問題への適用例として計算機実験を行なう．式 (26)〜式 (27)(⟨2.4.2⟩) は，出力層ユニットが複数ある一般的な場合としてベクトル表現しているが，関数近似問題では出力ユニットは 1 個となるためにベクトルではなくスカラーとなる．

ここで使用する関数は，関数最適化問題において，よく用いられている Schwefels 関数，Rastrigin 関数，および，Ridge 関数を関数近似問題の例として用いる．これらを表 6 に示し，それぞれの近似関数の概要を図 12〜図 14 に示す．

表 6: Learning Functions

(a)Schwefels Function $f(x,y) = -x\sin\sqrt{\lvert x\rvert} - y\sin\sqrt{\lvert y\rvert}$ $(-30.0 \le x \le 30.0, -30.0 \le y \le 30.0)$ $(-47.95 \le f(x,y) \le 47.95)$
(b) Rastrigin Function $f(x,y) = x^2 - 10\cos(2\pi x) + y^2 - 10\cos(2\pi y)$ $(-0.8 \le x \le 0.8, -0.8 \le y \le 0.8)$ $(-20.0 \le f(x,y) \le 20.5)$
(c)Ridge Function $f(x,y) = 2x^2 + 2xy + y^2$ $(-6.0 \le x \le 6.0, -6.0 \le y \le 6.0)$ $(0.0 \le f(x,y) \le 180.0)$

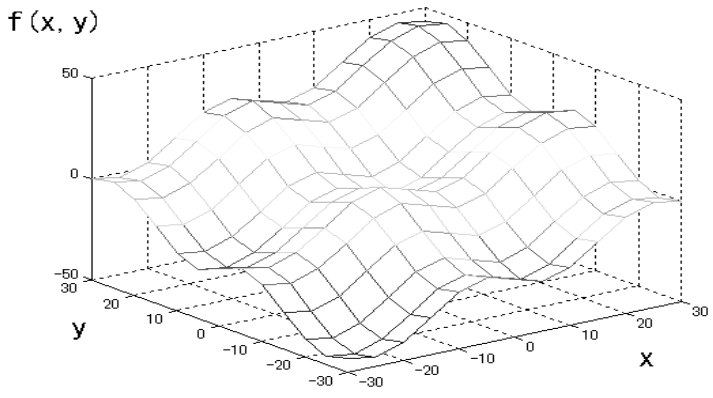

図 12: Schwefels function $f(x,y) = -x\sin\sqrt{|x|} - y\sin\sqrt{|y|}$.

関数最適化問題（3 次元）における Schwefels 関数の変数の範囲は，$(-512 \le x \le 512, -512 \le y \le 512)$ であり，$f(x,y) = f(420.968746, 420.968746) = 0$ を最適解に持つ．

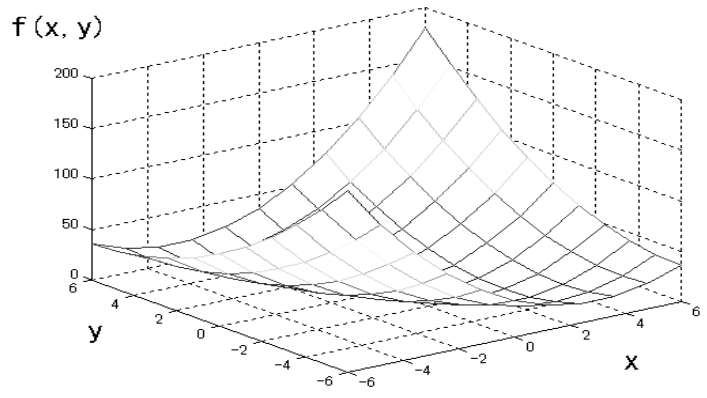

図 13: Ridge function $f(x,y) = 2x^2 + 2xy + y^2$.

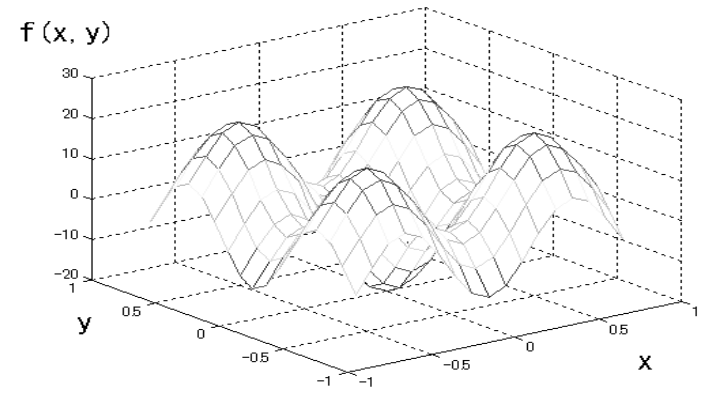

図 14: Rastrigin function $f(x,y) = x^2 - 10\cos(2\pi x) + y^2 - 10\cos(2\pi y)$.

また，Rastrigin 関数の変数の範囲は，($-5.12 \leq x \leq 5.12, -5.12 \leq y \leq 5.12$) であり，$f(x,y) = f(0,0) = 0$ を最適解に持つ．さらに，Ridge 関数の変数の範囲は，($-64 \leq x \leq 64$, $-64 \leq y \leq 64$) であり，$f(x,y) = f(0,0) = 0$ を最適解に持つ．

関数最適化問題と関数近似問題に対して，変数の範囲が異なる理由は，関数近似問題では，学習がどの手法を用いてもまったく進行しない場合があるために，範囲をせばめて学習を行なう．また，関数最適化問題では，変数の個数を自由に選択できるが，本論文では，x と y の 2 変数を扱っている．

4.2.1 教師データ

表 6(a)(b)(c) に対して，x,y 方向の刻み幅は，それぞれ，5.0，0.1，1.2 とし，領域を 13 × 13 (169 格子点)，17 × 17(289 格子点)，11 × 11(121 格子点) の格子に分割し，それらの格子点に対する値を教師データの集合とする．

基本的には，絶対値が大きいデータおよび入力パターン間の距離が近いにもかかわらず，それに対応する教師データ間の距離が大きいデータは，学習しにくい[25]．これらを重点的に学習するため学習データを学習しにくい順に並べる．これをほぼ均等に３グループに分け，第１段階では，第１グループを学習する．第２段階では，第２グループを追加し，第３段階では全データを学習する．

ここでは，多くの実験を行なったRastrigin関数を例に教師データの選択方法について詳しく述べる．

入力パターン間の距離が近いにもかかわらず，教師データ間の距離が大きく異なるデータの選択は，関数近似問題では偏微分値で代用できる．したがって，Rastrigin関数の教師データ選択には，x，y方向の偏微分値$f_x(x,y)$，$f_y(x,y)$を用いる．第1段階では，$|f(x,y)| \geq 0.90$を満たす教師データ24個，$|f_x(x,y)| \geq 60.3$を満たす教師データ34個，$|f_y(x,y)| \geq 60.3$を満たす教師データ34個が含まれ，そのうち2個が重複しているため，全体で90個（全データの31.1%）選択した．第2段階では，データ数は，58.1%（168個）選択した．これらのデータには，$|f(x,y)| \geq 0.80$を満たす教師データ52個，$|f_x(x,y)| \geq 60.1$を満たす教師データ68個，$|f_y(x,y)| \geq 60.1$を満たす教師データ68個が含まれ，そのうち20個が重複しているため，全体で168個となる．第3段階では，全教師データを使用する．

Schwefels関数およびRidge関数についても，同様な方針で教師データの選択を行なった．

4.2.2　比較対象手法と実験の諸設定

計算機実験では多段階学習法の有効性を検証するために，BP法，QPROP法，RPROP法，および多段階学習法の重み更新としてBP法，QPROP法，RPROP法の重み更新規則を採用した場合の６種類の比較実験を行なう．多段階学習法の実験では，第1段階から第3段階すべての段階で同一の重み更新規則を適用する．

NNの構成は，シグモイド素子を用い，入力層素子数2個，中間層素子数9個，出力層素子数1個からなるフィードフォワード型の3層構造のネットワークとする．中間層素子数は予備実験より決定した．さらに，重み係数の更新は，両手法とも2章で述べたように1エポック毎の一括更新方式を用いた．

QPROP法，RPROP法を多段階学習法の重み更新規則に用いる場合，**2.**で述べた重み更新式を用いて，各学習パターンについて重み更新量を求めた後，式(26)，(27)による誤差に応じた学習係数を用いて全パターンに対する$\Delta \boldsymbol{W}$を算出し，一括更新する．また，学習係数は，多段階学習法における基準の学習係数を$\eta = 0.8$とし，BP法単独の学習係数を0.8とした．QPROP法とRPROP法の各パラメータ値は，予備実験により決定した．詳しくは後述する．

学習結果を評価するRMSEに関しては，教師データに対する値を用い，重み係数の初期値を5通り設定した場合の平均値を用いる．重み係数の初期値は$[-0.01, 0.01]$の範囲でランダムに設定する．提案手法の学習回数は，1段階目2333エポック，2段階目2333エポック，3段階目2334エポックの合計7000エポックとし，BP法は7000エポックの学習で全教師データを常に用いる．実験に用いる計算機は，OS:Windows XP,CPU:Pentium 4, 3.0GHz, RAM:2GBである．

QPROP法に対する重み変更抑制係数$\lambda = 0.005$，最大変化量$\mu = 0.95$（重み係数の急激な変更による振動を避けるための最大変化量），$\rho = 0.80$とした．

さらに，RPROP 法に対しては，$\Delta_{\max} = 5.0$（上限），$\Delta_{\min} = 0.0025$（最小更新量），$\mu^+ = 0.97$（$m > 0$ のとき），$\mu^- = 0.61$（$m < 0$ のとき）とし実験を行なった．

ここで，QPROP 法および RPROP 法に対するパラメータは，多くの文献や多数の予備実験から決定した値である．

4.2.3 計算機実験結果 2 および考察

Ridge 関数および Rastrigin 関数，Schwefels 関数に対して，多段階学習法に BP 法，QPROP 法，RPROP 法を組み込んだ場合と，BP 法，QPROP 法，RPROP 法を単独で用いた場合の精度および学習時間を表 7〜表 10 に示す．表中の “+”印は，多段階学習法に組み込んだ手法を示し，数値は重みの初期値をランダムに設定した 5 回分の実験結果である．また，学習の精度は RMSE で評価している．

まず，学習の精度について述べる．表 7 および表 8 より，Ridge 関数と Rastrigin 関数に対しては，BP 法，QPROP 法，RPROP 法のそれぞれについて，“+”印付きの方が誤差は小さく，多段階学習法に組み込む効果が現われている．全体としては，+QPROP 法，+RPROP 法および RPROP 法の精度がよい．

表 9 より Schewefels 関数についても BP 法，QPROP 法に対しては，“+”印付きの方が誤差は小さくなっている．RPROP 法，+RPROP 法に対しては，平均 RMSE の値はそれぞれ 0.098，0.070 となっており，RPROP 法の方が多少精度がよい結果となっている．両者の最大誤差を見ると，RPROP 法の 0.103 に対して，+RPROP 法は 0.238 であり，この差が平均誤差に影響していると考えられる．

次に学習時間について述べる．表 10 より，Ridge 関数に対する QPROP 法と+QPROP 法の学習時間の平均はそれぞれ，624 秒，430 秒であり，多段階学習法を組み込むと，69% の学習時間になる．表 10 全体でも，組み込みによる学習時間は組み込まない場合と比較して，50% ∼ 69% となり，文献 (1) の s 段階学習における推定式 $\frac{s+1}{2s}$ に対して，$s = 3$ の場合の値にほぼ合致している．

表 7: RMSE for multi-stage learning and traditional methods in learning of Ridge function.

Function	Ridge		
Method	Mean	Max	Min
+BP	0.107	0.235	0.038
+QPROP	0.058	0.113	0.024
+RPROP	0.038	0.051	0.030
BP	0.241	0.243	0.234
QPROP	0.172	0.241	0.108
RPROP	0.050	0.080	0.023

表 8: RMSE for multi-stage learning and traditional methods in learning of Rastrigin function.

Function	Rastrigin		
Method	Mean	Max	Min
+BP	0.071	0.095	0.039
+QPROP	0.066	0.131	0.026
+RPROP	0.116	0.137	0.088
BP	0.259	0.483	0.193
QPROP	0.186	0.201	0.126
RPROP	0.254	0.618	0.091

表 9: RMSE for multi-stage learning and traditional methods in learning of Schwefels function.

Function	Schwefels		
Method	Mean	Max	Min
+BP	0.126	0.134	0.127
+QPROP	0.130	0.136	0.124
+RPROP	0.098	0.238	0.044
BP	0.295	0.343	0.244
QPROP	0.280	0.298	0.261
RPROP	0.070	0.103	0.053

表 10: Learning time for multi-stage learning and traditional methods in learning of three functions.

Learning	Learning Time(sec)		
Method	Ridge	Rastrigin	Schwefels
+BP	419	1204	574
+QPROP	430	1212	616
+RPROP	458	1218	634
BP	620	1956	966
QPROP	624	1984	977
RPROP	832	1992	1256

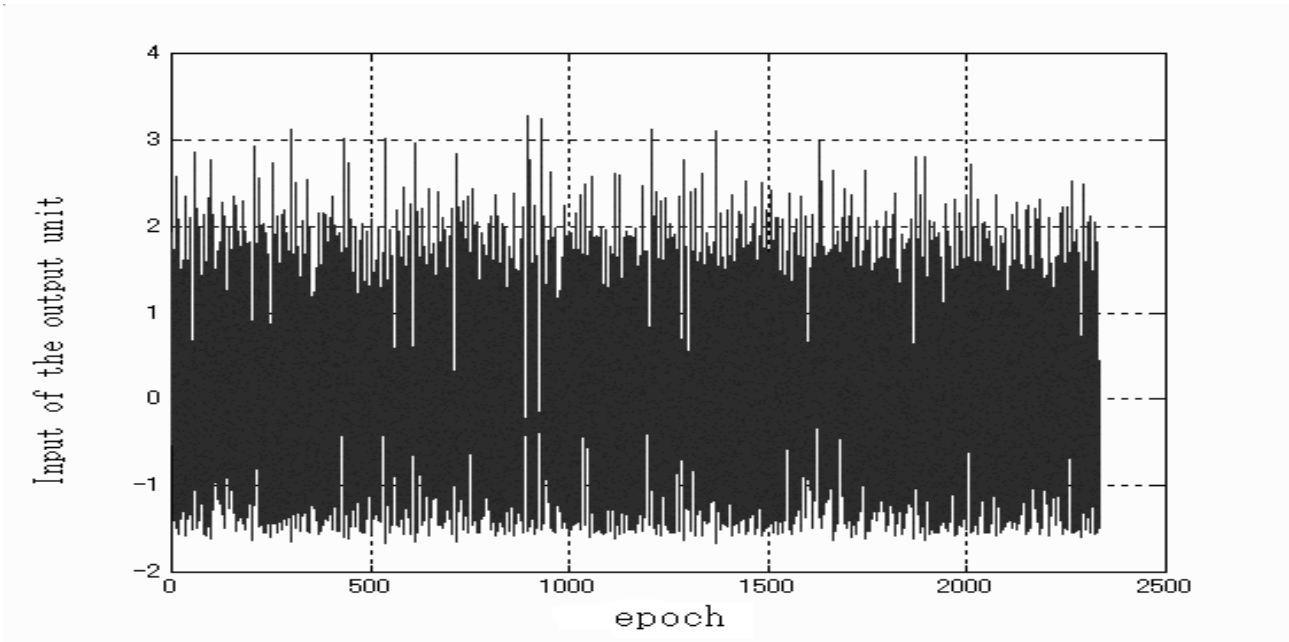

図 15: Input characteristics for the BP methods in the output unit (Rastrigin function, $\eta = 0.8$).

ここで学習時の第 3 段階における出力層素子への入力特性の典型例を図 15～ 図 18 に示す．これらの特性は, 学習開始時に学習データから任意の一つを選んで固定し, 各エポックにおいてこの学習データが入力されたときの, 出力層素子への入力, すなわち, 重みを介した中間層素子からの入力の総和を示した図である．横軸はエポック数を示し, 縦軸は出力層素子への入力値である．また，横軸のエポック数は第 2 段階の最終エポックを 0 として示している．

図 15～ 図 18 に示す特性に対する学習終了後の RMSE は，それぞれ 0.206，0.039，0.024，0.047 である．図 15 は，Rastrigin 関数に対する BP 法の特性である．多段階学習法ではないが, 横軸のエポック数は第 2 段階の最終エポックに相当するエポックを 0 として示している．不規則で大きな振幅の振動が最後まで継続する特性であり, 学習後の RMSE 値は 0.206 と大きく, 学習が失敗した例である．図 16 は，Rastrigin 関数に対する+BP 法の特性である．規則振動に不規則振動が加わった振動現象を継続している．図 17 は，Ridge 関数に対する+QPROP 法の特性である．図 18 は，Ridge 関数に対する+RPROP 法の特性である．240 エポックを境に，非振動状態から規則振動に移っている．図 15～ 図 18 に対するフーリエ変換（Fourier Transform）を行なった結果をそれぞれ図 19～ 図 22 に示す．縦軸は振幅「Magnitude」，横軸の名称は周波数「frequency」で, 単位は「frequency/epoch」である．横軸は，2334 エポックの学習を行なった出力層素子への入力特性をフーリエ変換したものである．したがって，0.5 における振幅成分が最大振動数となり，最大周波数は，1167 振動 (2334 エポックの 0.5 倍) であり，学習中常に振動をしている成分である．

周波数に対する振幅の大きさの分布は異なるが，共に低周波から高周波の振幅を広く含む結果となっている．RPROP 法を用いた図 22 に対する振幅は，振動を考慮した重み更新を行なっているために他と比較して非常に小さくなっている．

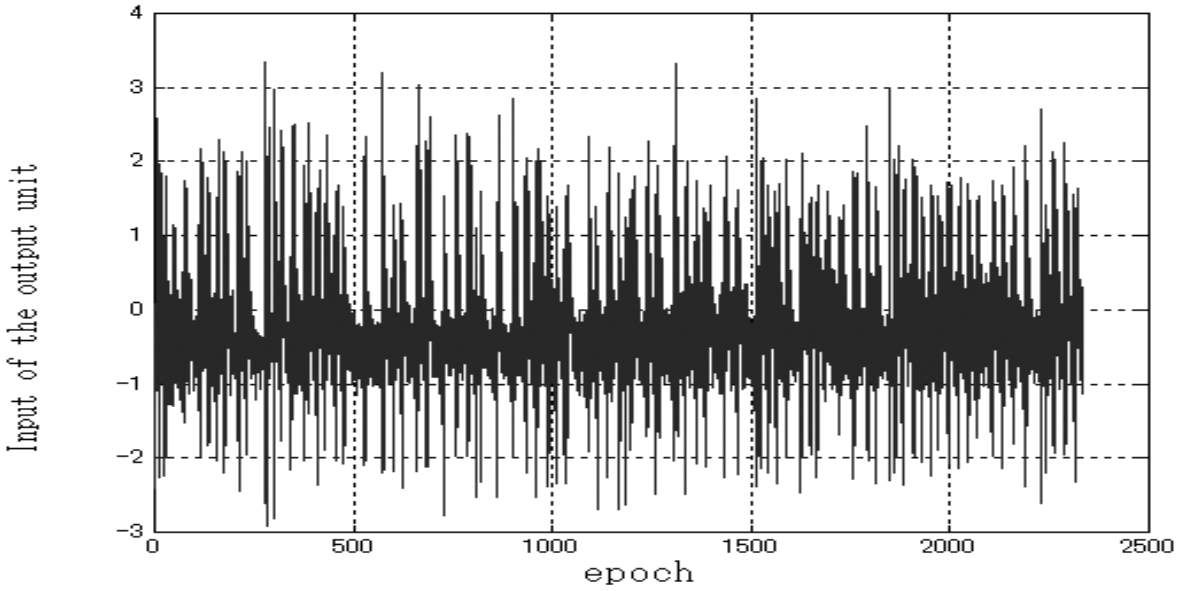

図 16: Input characteristics for the +BP methods in the output unit (Rastrigin function, $\eta = 0.8$).

4.2.4 BP 法と振動の有効性

BP 法において学習率を小さな値にすると, 図 15 のように振動することは同様であるが, 振幅は小さくなる．さらに学習率を小さくすると振動は止まるが, 振動の有無にかかわらず誤差は大きく, Ridge 関数，Rastrigin 関数，Schwefels 関数に対する良好な学習ができない.

一方, 多段階学習においても学習率 η (式 (26) 参照) を減少させると, BP 法と同様に振動現象は生じなくなるが, 良好な学習はできない．例えば, Rastrigin 関数の学習に学習率 $\eta = 0.05$ で＋ BP 法を適用した場合, 学習中に振動は生じないが RMSE 値は 0.135 となり誤差は大きい．これに対して, 学習率 $\eta = 0.80$ を用いた学習では振動現象が生じ, 表 8 より RMSE 値は 0.071 である．計算機実験では, 学習率 $\eta = 0.2 \sim 0.9$ の範囲で RMSE 値が 0.1 以下になることを確認した．Rastrigin 関数に対し，学習率を $\eta = 0.05 \sim 0.9$ の範囲で変化させた場合の BP 法および+BP 法に対する RMSE 値と振動の有無に対する結果を表 11 に示す.

以上のことより, 学習対象が複雑になると, 学習には振動現象が重要な役割を果たすといえる．したがって, 多段階学習法では振動を抑制する効果がある慣性項を取り入れていない [25]．また, 学習中の振動は外部から与えるのではなく, 自発的に生じる．振動現象が自発的に生じる原因については **4.2.5** で考察する.

4.2.5 振動現象発生のメカニズムと学習中の挙動

振動状態と非振動状態

図 15～ 図 18 からわかるように，出力層素子への入力特性に生じる振動現象にはさまざまなタイプがある．図 17 では，1900 エポック付近で振動状態から非振動状態になる場合

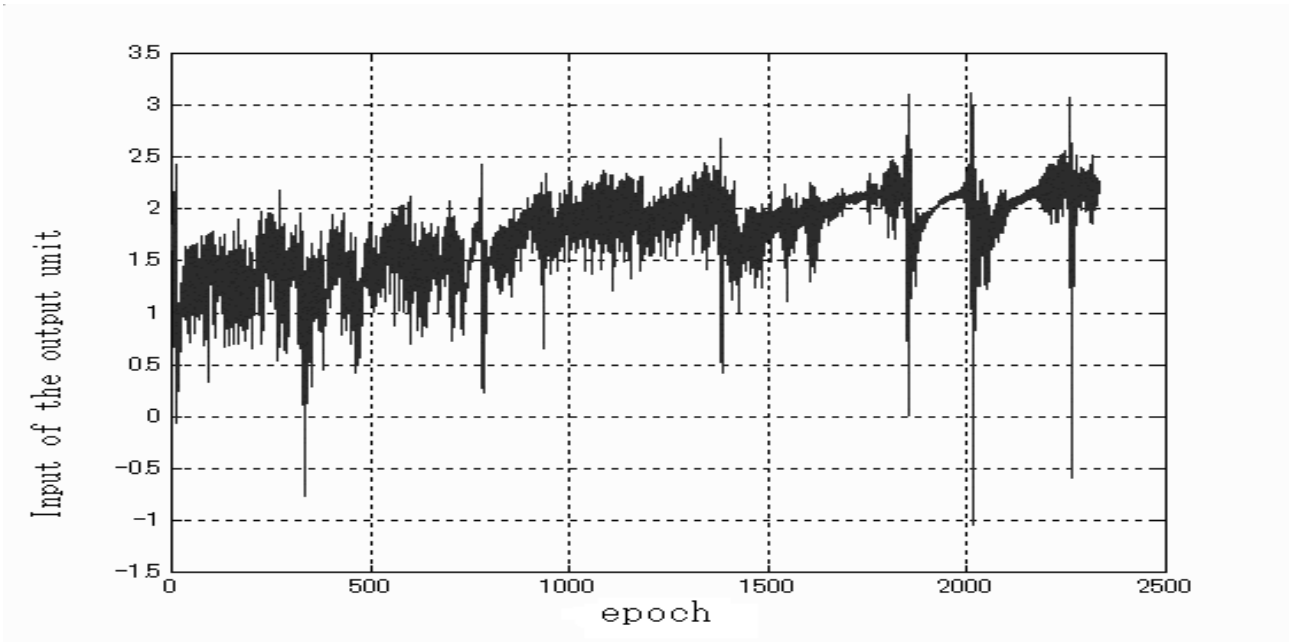

図 17: Input characteristics for the +QPROP methods in the output unit (Ridge function , $\eta = 0.8$).

や，非振動状態から振動状態になる場合がある．また，図 18 の 250 エポック付近は，非振動状態から振動状態になる場合である．振動状態から非振動状態および非振動状態から振動状態に変化する時の $\Delta \boldsymbol{W}^{+}(t)$(式 (32) 参照)，および，$\Delta \boldsymbol{W}^{-}(t)$(式 (33) 参照) の大きさや方向の典型的な変化の概略を図 23 に示す．合成ベクトル $\Delta \boldsymbol{W}(t)$(式 (34) 参照) の向きは常に右向きを基準として描いている．振動状態は，$\frac{|\Delta \boldsymbol{W}^{+}(t)|}{|\Delta \boldsymbol{W}^{-}(t)|} < 1$ と $\frac{|\Delta \boldsymbol{W}^{+}(t)|}{|\Delta \boldsymbol{W}^{-}(t)|} > 1$ の繰り返しとなり，非振動状態で学習が安定している場合には全体の誤差が単調減少となり，

$$\frac{|\Delta \boldsymbol{W}^{+}(t)|}{|\Delta \boldsymbol{W}^{-}(t)|} < 1 \tag{49}$$

表 11: RMSE for multi-stage learning and traditional methods by using Rastrigin function.

Function	Rastrigin +BP		Rastrigin BP	
Learnng coefficient	Mean	Oscillation	Mean	Oscillation
0.05	0.135		0.035	○
0.10	0.138	○	0.025	○
0.20	0.096	○	0.045	○
0.30	0.081	○	0.162	○
0.40	0.039	○	0.183	○
0.50	0.026	○	0.199	○
0.60	0.027	○	0.185	○
0.70	0.037	○	0.203	○
0.80	0.071	○	0.233	○
0.90	0.050	○	0.618	○

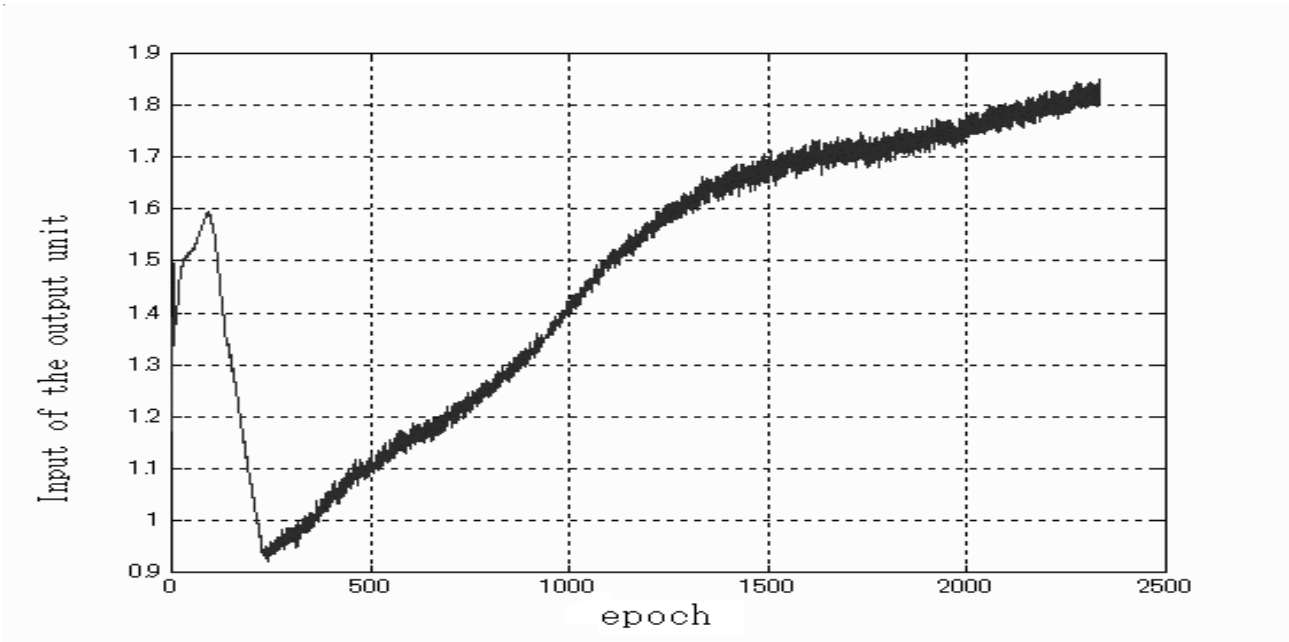

図 18: Input characteristics for the +RPROP methods in the output unit (Ridge function, $\eta = 0.8$).

となっていることが特徴で，常に誤差減少ベクトルが有効に働いている．また，非振動状態の時，誤差増加パターン数 $|\boldsymbol{P}^+(t)|$ と誤差減少パターン数 $|\boldsymbol{P}^-(t)|$ は，エポックが進んでもほとんど変化せず，$\frac{|\boldsymbol{P}^+(t)|}{|\boldsymbol{P}^-(t)|} \leq 1$ の関係が継続する．図 23 における $a1$ は，

$$|\Delta \boldsymbol{W}^+(t)| > |\Delta \boldsymbol{W}^-(t)| \tag{50}$$

の場合を示し，逆に，a2 は，

$$|\Delta \boldsymbol{W}^-(t)| > |\Delta \boldsymbol{W}^+(t)| \tag{51}$$

の場合であり，a3, a4 は，a1, a2 に対し，$|\Delta \boldsymbol{W}(t)|$ が小さいことを示している．$a5$ では，$|\Delta \boldsymbol{W}^+(t)| \approx 0$ で $|\Delta \boldsymbol{W}^-(t)|$ と $|\Delta \boldsymbol{W}(t)|$ はほぼ一致する．これは，振動していない場合は，合成ベクトル $\Delta \boldsymbol{W}(t)$ は，減少ベクトルが支配的であり，大きさは小さく，この状態が持続することを意味する．非振動状態では，$\frac{|\Delta \boldsymbol{W}^+(t)|}{|\Delta \boldsymbol{W}^-(t)|} \ll 1$ の関係を保ち，全体の誤差は単調に減少する．

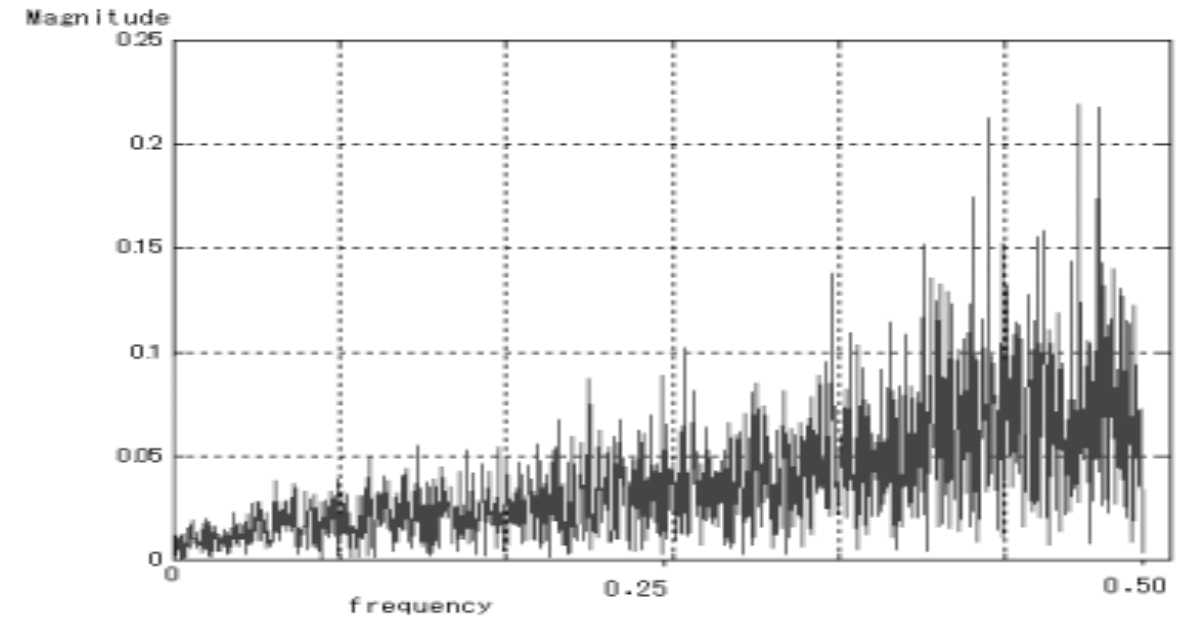

図 19: Frequency spectrum of input characteristics for the BP methods in the output unit (Rastrigin function, $\eta = 0.8$).

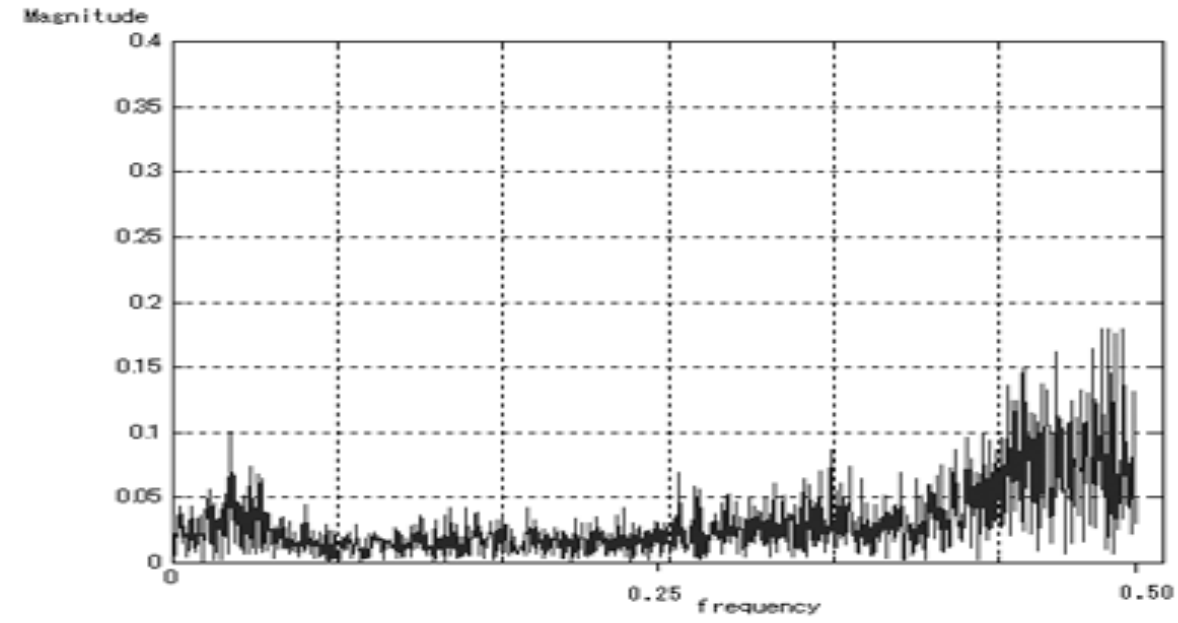

図 20: Frequency spectrum of input characteristics for the +BP methods in the output unit (Rastrigin function, $\eta = 0.8$).

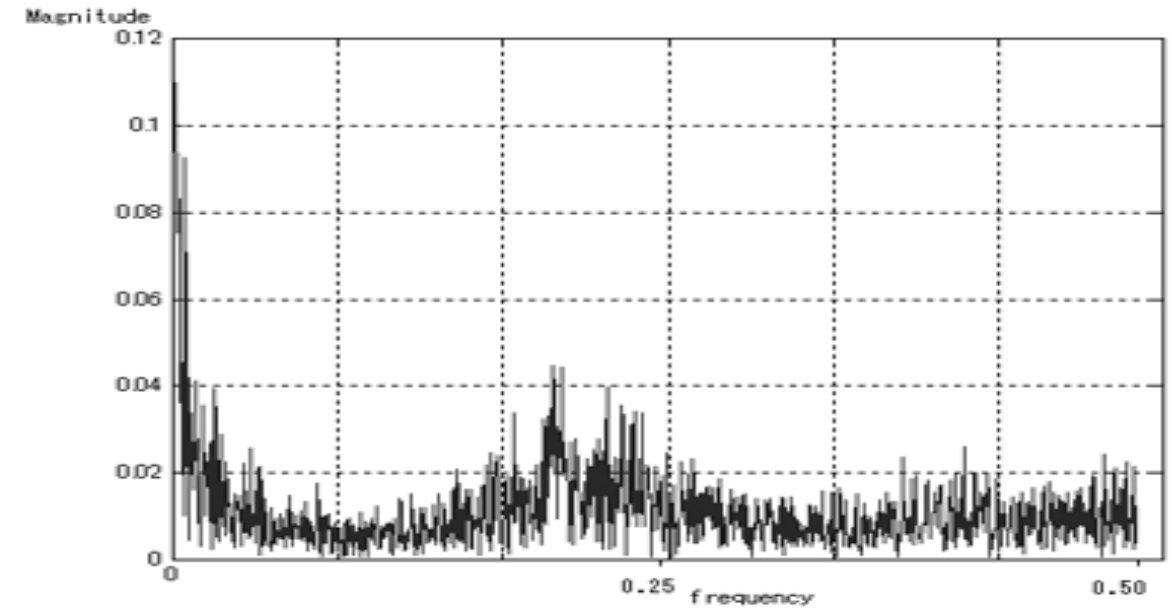

図 21: Frequency spectrum of input characteristics for the +QPROP methods in the output unit (Ridge function, $\eta = 0.8$).

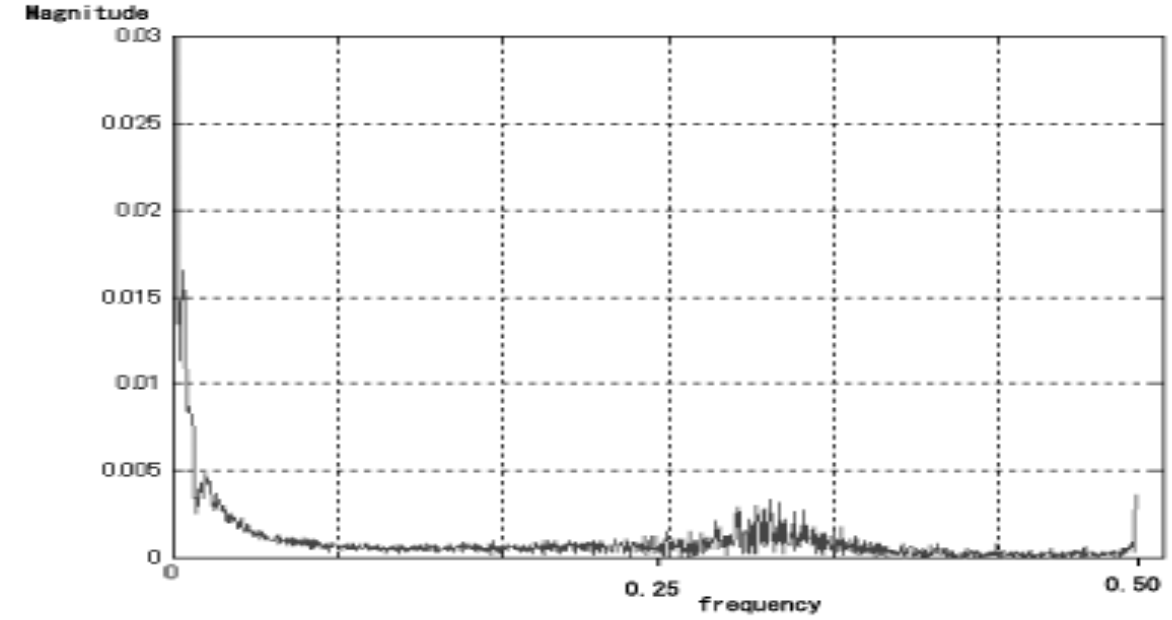

図 22: Frequency spectrum of input characteristics for the +RPROP methods in the output unit (Ridge function, $\eta = 0.8$).

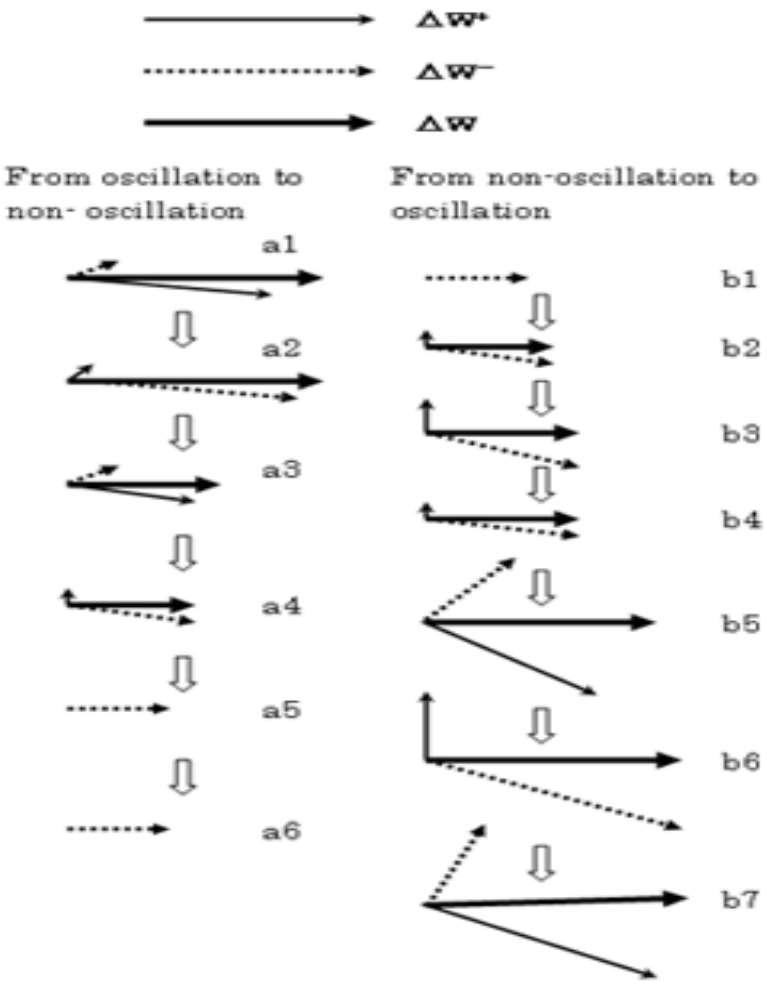

図 23: The weight renewal vector for the multi-stage learning methods.

振動の発生

図 23 の b1∼b7 は，非振動から振動に移るときのベクトル図である．非振動状態において，すべてのパターンについて，誤差が減少する方向に $\Delta \boldsymbol{W}$ が更新されていたとしても，学習の最後までその状態が続くとは一般的に考えられないので，あるエポック t で，あるパターン $p \in \boldsymbol{P}^+(t)$ が生じたとする．他のパターンは $\boldsymbol{P}^-(t)$ に属するので，重み更新量 $|\Delta \boldsymbol{W}(t)|$ は，$|\Delta \boldsymbol{W}^-(t)|$ が支配的である．そのため，重みはパターン p に対する誤差を増加させる方向に修正される．したがって，

$$|\Delta \boldsymbol{W}^+(t+1)| > |\Delta \boldsymbol{W}^+(t)| \tag{52}$$

の状態が何エポックか継続する．式 (26) から明らかなように，学習率は誤差の二乗で効くため，$|\Delta \boldsymbol{W}^+(t)|$ の増加速度は加速する．それにより，エポック t で $p \in \boldsymbol{P}^+(t)$ だったパターン p は，あるエポック $t'(> t)$ において $p \in \boldsymbol{P}^-(t')$ となる．ただし，エポック $(t'+1)$ ではパターン p の誤差は小さくなるので，再び $p \in \boldsymbol{P}^+(t'+1)$ となり，このことが繰り返されて，振動現象が生じると考えられる．

学習後に誤差が大きい学習データの学習中の挙動

学習パターンの中には，学習終了後にも誤差が大きいままのパターンが存在する．そのようなパターンは学習方法や学習ごとに変わるのではなく，特定のパターンに固定される．例えば，Rastrigin 関数の学習では，$(x, y) = (1, 1), (1, 15), (1, 17)$ が該当する．ここでは，それらのうち，$p^* \equiv (1, 1)$ を例として，パターン p^* の誤差がなぜ学習後でも小さくならないかを考察する．

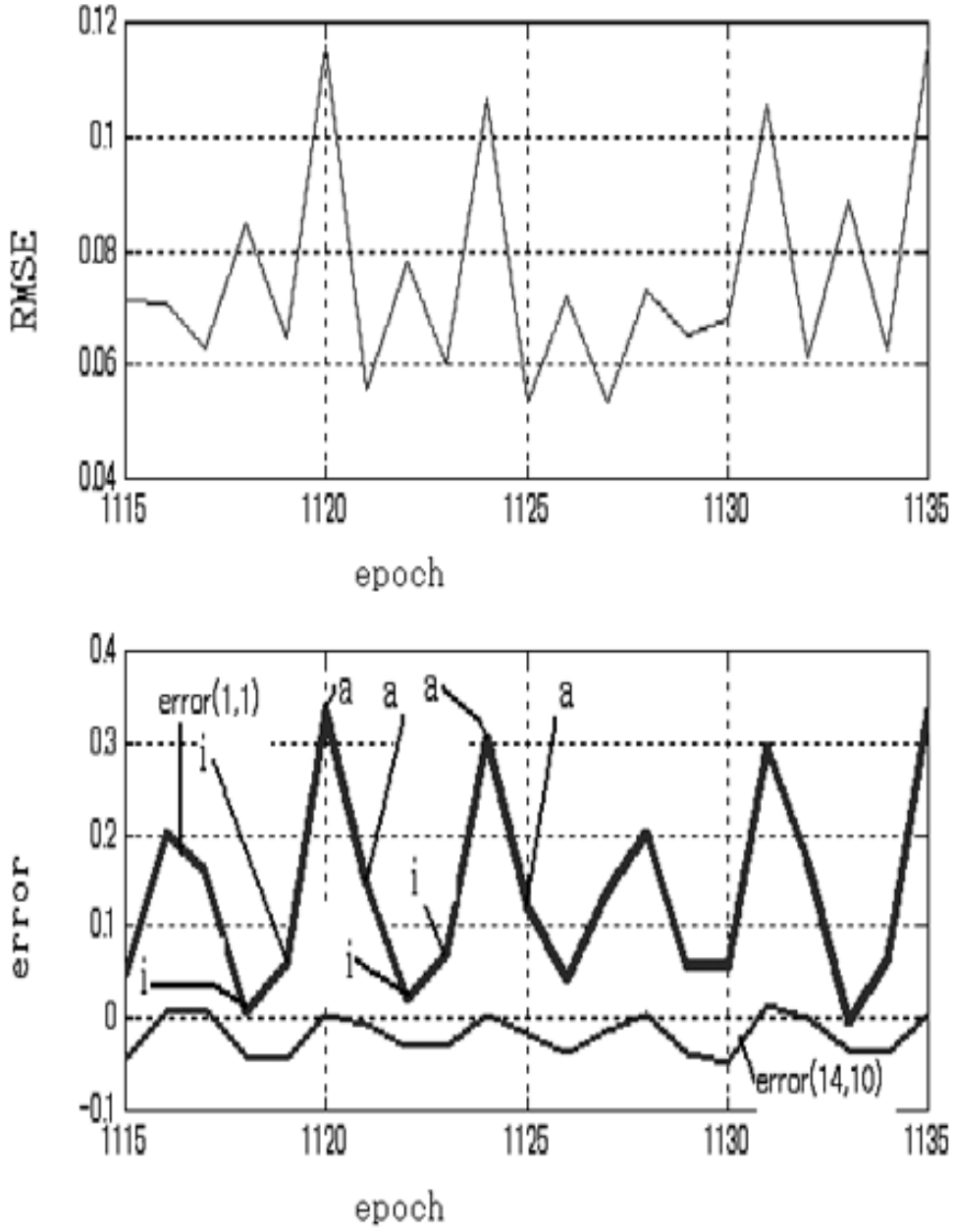

図 24: The behavior of the paturn p^* between 1115∼1135 epochs at the third stage in the multi-stage learning($\eta = 0.4$).

式 (28)，(29) における $\boldsymbol{P}^{\pm}(t)$ に関して，

$$|\boldsymbol{P}^{\pm}(t)| > |\boldsymbol{P}^{\mp}(t)| \text{（複合同順）} \tag{53}$$

のとき，前者を多数派，後者を少数派と呼ぶことにする．一般的に，重みの修正分に関して，

$$|\boldsymbol{P}^{\pm}(t)| > |\boldsymbol{P}^{\mp}(t)| \tag{54}$$

ならば

$$|\Delta\boldsymbol{W}^{\pm}(t)| > |\Delta\boldsymbol{W}^{\mp}(t)| \text{（複合同順）} \tag{55}$$

の場合がほとんどである．$|\Delta\boldsymbol{W}^{-}(t)| > |\Delta\boldsymbol{W}^{+}(t)|$ のとき，RMSE 値は減少する．

パターン p^* は教師データの絶対値が大きいデータであり, 多段階学習における第 1 段階から学習に用いられるデータである．例えば, Rastrigin 関数の学習で+BP 法を用いたある実験において, パターン p^* の誤差の絶対値の初期値は $\epsilon_{rr}^{p^*}(0) = 0.164$ であり, 第 1 段階終了時では 0.013∼0.014 の範囲で振動, 第 2 段階終了時では 0.020∼0.028 の範囲で振動, 第 3 段階終了時では 0.094∼0.125 の範囲で振動していた。このことから, 第 1 段階では学習しにくいと判断されたデータだけを学習するので, ある程度誤差は小さくなるが, 第 2, 第 3 段階で新たな学習データが加わると, それらデータの初期の誤差は大きいので, 重み修正ベクトルは新たに加わったデータの誤差を減少させる方向に引き摺られる．したがって, パターン p^* は少数派に属することになり, 一度減少した誤差が大きくなる．それに対して, 第 3 段階で加わった学習データ, 例えばパターン $p^\dagger \equiv$(14,10) に関して, 誤差の絶対値の初期値（乱数により決定）は $\epsilon_{rr}^{p^\dagger}(4667) = 0.029$ であるが, 5 エポック程度の学習により, 0.011∼0.047 の範囲で振動するようになり, パターン p^* と比較すると小さい．この時点でパターン p^* とパターン $p^\dagger$ の誤差の大小に関する相対的な関係が固定される．以後学習終了時まで，この相対的な関係が維持される．

第 3 段階の 1115〜1135 エポックにおける RMSE 値とパターン p^*, $p^\dagger$ の挙動を図 24 に示す．上図は RMSE 値, 下図は誤差を示し, 太線がパターン p^*, 細線が $p^\dagger$ に対応している．“a” は多数派 (majority), “i” は少数派 (minority) に属していることを示す．図 24 からわかるように, パターン p^* は少数派, 少数派, 多数派, 多数派の繰り返しになっていることが多く, 多数派の時に誤差は減少し, 少数派になると誤差は増加する．$p^\dagger$ も同様であり, 相対的な誤差の大小関係を保ったまま, 誤差の平均はほとんど減少しない状態が学習終了時まで継続する以上のことから, パターン p^* のような, 第 1 段階で一度誤差が減少してもその後増加し, 学習が終了しても誤差が大きい学習データが生じることになる．

5　あとがき

本論文では，特に，学習が困難と言われている関数学習において，最も基本的な，しかも，実用的には非常に多く使用されている BP 法に対し，以下に述べる 3 個の工夫を施した結果，学習時に効果的な自発的振動が発生し，従来から学習が困難とされていた関数学習に対して学習が行われ，学習精度および学習時間の向上を行なうことができた．さらに，出力層素子への入力特性の振動に対し，個々の学習データの誤差の大きさと振動波形の発生，振動継続条件を，新たに重み更新ベクトルを導入し，考察を行なった．

実際に，NN を用いた学習において，学習する前に教師データを学習のしやすさに着目した分類，その分類に基づく多段階学習，誤差の大きさに応じた学習係数の動的調整を特徴とした総合的な学習方法を提案した．さらに，出力層素子への入力特性に着目し，学習の継続の可否を判定する振幅減少条件，目標値捕捉条件を提案した．そして，提案手法を検証するために関数近似問題を対象とした計算機実験を行なった．その結果，従来の方法では，学習が困難な対象に対して，提案手法では学習が容易に行われ，学習時間も短縮される結果となった．したがって，従来からシグモイド素子から構成される NN では困難とされている学習を可能とし，その有効性を示した．また，多段階学習を取り入れることによる学習時間の短縮効果，振動現象の利用に対する有効性も示した．提案手法は特別な素子を用いるわけではなく，一般的なシグモイド素子だけを用いている．したがって，関数近似問題以外の問題に対しても容易に適用可能であると考えられる．また，時系列パターンを用いた予測・推定問題等における学習に対し，学習しにくいデータを多く含む問題にも有効に利用できると考えられる．さらに，本論文では最も基本的な BP 法に基づく効率的学習法を提案したが，BP 法の拡張である QP 法や弾力性 BP などにも適用可能で，具体的には, 多段階学習法の学習則として QPROP 法と RPROP 法を用いた場合にも, 学習精度および学習時間の両方が改善されることを計算機実験により確認した．また, 学習においては学習曲線の振動現象が学習性能に重要な役割を果たすと共に, 振動現象の発生メカニズムについて考察した．さらに, 学習が終了しても誤差が小さくならないデータに対して, 学習中の挙動を観察することにより, その理由を検討した.

多段階学習法においては, 学習データを分類する際, 入力ベクトル間の距離に対する出力ベクトル間の距離や出力ベクトルの大きさ (本論文で扱った関数近似問題では出力値) の関係に着目している．今後の課題として, このような距離の関係を利用できない, 例えば判別問題のような学習対象に対しても適用できるように, 多段階学習法をさらに拡張することが挙げられる.

最後に本研究において，公表した我々の論文に対して，それを参考として 2 個の発表論文が公表されていることを報告する．論文 [13] に対して，ニューロユニットの出力関数に三角多項式を組み込んだ論文 [40] がある．これは，医用関係に応用されている．さらに，敬

愛大学国際研究の第 12 号の文献を参考にした論文 [41] は，電気学会の研究会で使用され発表されている．工作機械の精密制御に応用され，研究会で口頭発表されている．

前項で述べたように，学習データの性質によって振動の様子が決定されている．すなわち，誤差が常に少ない学習データのグループ，最終的に誤差が常に大きくなる学習データのグループが存在する．学習しやすい学習データのグループ，すなわち，誤差が常に少ないグループは，必ず多数派に属している．一時的には誤差が少なくなっても最終的に誤差が大きくなる学習データのグループは，前節で述べたように，少数派, 少数派, 多数派, 多数派のように規則的な繰り返しで比較的周期的となり，最終的には誤差は大きくなる．

しかし，学習が困難な関数学習において，誤差の少ない学習が行われる時，任意の出力層素子への振動曲線は，複雑な振動特性になる．理由についてはまだ検討の余地がある．

任意の学習データに対して，学習しやすいデータと学習しにくいデータの中間的な学習データの存在が影響していると考えられる．データとしてはここでは述べないが，時折，ほとんど多数派に属している時が多いにもかかわらず，時折，少数派になり，周期的に変化しない学習データが存在する．また，逆に，比較的学習しにくいデータであったものが，学習しやすいデータに変化する場合（非周期的）があり，不安定である学習データが存在する．このデータが振動曲線をより複雑にし，誤差に関しては，良好な結果を生じさせているのではないかと考察できる．

最後に，複素ニューラルネットワークを用い，ユニット数を削減し，実関数の学習をすることが試みられていることを，紹介する（論文 [43]）．ここでは，複素 BP 法を用いた場合は，学習が進行しないこと，及び，複素 BFGS 法や複素 SS 法を用いた場合には，誤差が少なくなったことが述べられている．構成の容易な複素 BP 法では成功しない問題ということで，興味を引き起こす今後の研究課題が残されたものと考えられる．

参考文献

[1] 元木誠，小圷成一，平田廣則:“パルスネットワークのための入出力パルスのタイミングを調節する教師あり学習則”, 信学論, (D-II), Vol.J89-D-II, No.12, pp.2744-2756 (2006).

[2] 新田徹:“複素ニューラルネットワークにおける一意性定理とパラメーターの冗長性” , 信学論, (D-II), Vol.J85, No.5, pp.796-804 (2002)

[3] 幸田憲明，松井伸之，西村治彦:“量子ビットニューロンモデルによる階層型ネットワーク” , 信学論, (D-II), Vol.J85-D-II, No.4, pp.641-648 (2002).

[4] M.Riedmiller and H.Braun:“A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm”, Proc. ICNN, San Fransisco (1993).

[5] Fahlman,S.E.:“An Empirical Study of Learning Speed in Back-Propagation Network”, Technical Report CMU-CS-88-162, Carnegie-Mellon University, Computer Science Dept., (1988).

[6] 原一之，中山謙二:“汎化能力獲得のためのデータ選択法と学習法”, 信学技報, NC98-35, pp.39-46 (1998).

[7] 山川宏，増本大器，木本隆，長田茂美:“逐次学習における学習データの選択と追加的出力補正”, 信学技報, NC-92-99, pp.33-40 (1993).

[8] C.Cachin:“Pedagogical pattern selection strategies”,Neural Networks Vol.7, No.1, pp.175-181 (1994).

[9] 小原和博，中村行宏:“バックプロパゲーション・ニューラルネットワークへの学習セットの選択的提示法”, 電学論 C, Vol.117, No.9, pp.1281-1290 (1997)

[10] 林秀樹，岡部洋一:“ニューラルネットワークにおける振動生成のメカニズム” , 信学技報, NC-96-22, pp.1-8 (1996).

[11] 金丸隆志，関根優年:“ニューラルネットワークの canonical model がみせる振動同期現象”, 信学技報, NC2003-138, pp.17-22 (2004).

[12] 松井伸之，石見憲一:“しきい値ゆらぎをもつニューロンモデルを用いた階層型ニューラルネットワーク”, 電学論 C, Vol.114-C, No.11, pp.1208-1213 (1994).

[13] 田口功，須貝康雄:“出力層素子の入力特性とそれに基づくニューラルネットワークの学習の効率化”, 電気学会電子．情報．システム部門講演論文集, pp.931-934 (2004).

[14] Chong-Ho Choi and Jin Young Choi，“Piecewise Interpolation Capabilities for Function Approximation”, IEEE Trans. on Neural Networks, Vol.5, No.6, pp.936-944 (1994).

[15] Ting Wang，須貝康雄:“非線形多変数関数近似のためのウェーブレットニューラルネットワーク”, 電学論 C, Vol.120-C, No.2, pp.185-193 (2000).

[16] L. K.Jones:“Constructive Approximations for Neural Networks by Sigmoidal Functions”, Proc. IEEE, Vol.78, No.10, pp.(1990).

[17] Ting Wang and Yassuo Sugai:“A Wavelet Neural Network for the Approximation of Nonlinear Multivariable Functions”, Proc.of IEEE International Conference on System, Man, and Cybernetics, III, pp.378-383 (1999).

[18] B.Irie and S.Miyake:“Capabilities of Three Layered Perceptrons”, Proc.ICNN, Vol.1, pp.641-648 (1988).

[19] K. Funahashi:“On the Approximate Realization of Continuous Mapping by Neural Networks”, Vol.2, No.3, pp.183-192 (1989).

[20] Charles K.Chui:“An Introduction to Wavelets”, Academic Press (1992).

[21] 武智宏親，村上研二:“学習係数の動的制御によるニューラルネットワークの動作特性”, 信学技報, NC-93-76, pp.65-71 (1994).

[22] 猪飼武夫，山崎秀聡，小迫秀夫:“逆伝搬学習法における動的学習率の適応的決定法”, 信学技報, NC-90-67, pp.95-101 (1992).

[23] 須貝康雄，堀部浩，川瀬太郎:“基準需要を利用したニューラルネットによる翌日最大電力需要予測”, 電学論B, Vol.117-B, No.6, pp.872-879 (1997).

[24] 梅原宗一，山崎輝，須貝康雄:“サポートベクタマシーンとニューラルネットワークに基づく降水量推定システム”, 信学論 (D-II), Vol.J86-D-II, No.7, pp.1090-1098 (2003).

[25] 田口功，須貝康雄:“教師データの選択と出力層素子への入力特性に基づくニューラルネットワークの効率的学習法”, 電学論 C, Vol.129-C, No.4, pp.1208-1213 (2009).

[26] M. Riedmiller and H. Braun:“A DirectbAdaptive Method for Faster Backpropagation Learning: The RPROP Algorithm”, Proc. ICNN, San Fransisco, (1993).

[27] D.E.Rumelhart, J.L.McCleland, and the PDP Research Group: “Parallel Distributed Processing Vo.1”, MIT Press (1986).

[28] McCulloch, W.S., and Pitts, W.:“A logical calculus of the ideas immanent in nervous activity,” Bulletin of Math., 5, pp.115-133 (1943).

[29] 八名和男監訳:“ニューラルコンピューティング入門”, 海文堂 ，pp77〜pp79，(1993).

[30] 熊沢逸夫:学習とニューラルネットワーク，森北出版，東京，5 月，1997.

[31] 萩原将文:ニューロ・ファジイ・遺伝的アルゴリズム，産業図書，pp.2124，1994.

[32] 廣瀬　明:“複素ニューラルネットワーク”, サイエンス社 ，pp55〜pp59，(1993).

[33] 新田　徹:“回転を学習した複素 BP ネットワークのふるまい”, 情報処理学会論文誌, Vol.34, No.1, pp.39-51 (1993).

[34] Amari, S:A Theory of adaptive pattern classifiers, IEEETrans., EC 16-3, pp.299-307(1969).

[35] 甘利俊一:神経回路網の数理, 産業図書 (1978).

[36] 福島邦彦:神経回路と自己組織化, 共立出版,(1979).

[37] 中野:アソシアトロン, 昭晃堂 (1979).

[38] kohonen, T.:self-Organization and Associative Memory, 2nd ed., Springer(1988).

[39] 松葉育雄:ニューラルシステムによる情報処理, 昭晃堂 (1993).

[40] 池田明日美, 吉村宏紀, 堀磨伊也, 清水忠昭, 岩井儀雄, 岸田　悟: “TPUnit ニューラルネットワークを用いた胸部 X 線画像の異常陰影検出システムの提案”, 電気・情報関連学会中国支部連合大会講演論文集, pp.456-457 (2012), Oct.

[41] 立田昌也, 呉　世訓, 堀　洋一:“工作機械の精密制御のための最適パラメータ探索法”, 電気学会研究会資料.　C/ 電気学会産業計測制御研究会［編］, pp.61-65 (2007).

[42] 渡辺真也，榊原一紀:“単目的最適化問題における多目的化とその有効性”, 情報処理学会論文誌, Vol.46, No.4, pp.1-10 (2005).

[43] 佐藤聖也，中野良平:“出力が実数の問題に特化した目的関数を用いた複素多層パーセプトロン探索”, 計測自動制御学会　システム・情報部門学術講演会 2013, 2013 年 11 月 18 日〜20 日, (2013).

著者関連論文リスト

論文誌 (査読あり)

1) 田口　功・須貝康雄：教師データの選択と出力層素子への入力特性に基づくニューラルネットワークの効率的学習法，電学論，Vol.129，NO.4，pp.726-734，2009 年.

2)Isao Taguchi･Yasuo Sugai: Oscillation Behavior for the Layered Neural Networks Based on the Selection of Training Data by Using Rastrigin Function, International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 4, pp.21-27, April 2013.

3)Isao Taguchi･Yasuo Sugai:A Multi-stage Learning Method with the Selection of Training Data for the Layered Neural Networks, International journal of Scientific and Research Publication, Volume 3, Issue 4, pp.1-6, April 2013.

論文誌 (査読中論文)

4)Isao Taguchi・Yasuo Sugai:The Extended Multi-stage Learning Using Voluntary Oscillation and Its Learning Performance,International Journal of Computer and Information Engineering, WASET JOURNAL PAPER.

国際会議 (査読あり)

1)Isao Taguchi・Yasuo Sugai:Oscillation Effect of the Multi-stage Learning Method for Layered Neural Networks and Its Analysis, WORLD ACADEMY OF SCIENCE ENGINEERING AND TECHNOLOGY，pp.843-848, SEPTEMBER 28-30, 2011.

国内講演会 (査読なし)

1) 田口　功・須貝康雄：出力層素子の入力特性とそれに基づくニューラルネットワークの学習の効率化，電気学会　電子・情報・システム部門，GS12-2，宇都宮大学，2004 年 9 月.

2) 田口　功・須貝康雄：出力層素子の入力特性に基づく多層ニューラルネットワークの多段階学習法，電気学会　電子・情報・システム部門，GS5-6，公立函館未来大学，pp734-737，2008 年 8 月.

その他

1) 田口　功：ニューラルネットワークの関数近似におけるシグモイド関数の問題点，敬愛大学国際研究，第６号，pp.77-90，2000 年 11 月.

2) 田口　功・須貝康雄：３層ニューラルネットワークの関数学習における誤差領域，敬愛大学国際研究，第 10 号，pp.63-93，2002 年 11 月.

3) 田口　功・須貝康雄：３層ニューラルネットワークにおける誤差特性を利用した関数学習の加速化，敬愛大学国際研究，第 13 号，pp.87-98，2004 年 6 月.

4)Isao Taguchi・Yasuo Sugai:An Efficient Learning Method for Layered Neural Networks Based on Selection of Training Data and Input Characteristics of an Output Layer Unit, Wiley Periodicals,Inc，pp.57-67, 2012.

謝辞

本研究を行なうに当たり，非常に幅広く直接ご指導いただいた工学研究科の須貝康雄教授に心から感謝いたします．特に，学習が困難となる関数の存在に対しては，今考えてみると，その効果が非常に大きかったと思う．2，3 日パソコンを走らせ学習が全然行われなかったこととパソコンが壊れ，マザーボードを交換したことは，今でも頭に浮かぶことである．研究進展に対し，要所要所でその糸口を与えてくれたこと，細かな研究の進め方の助言，論文の数式化の方法，修士課程とは異なる研究の細かな指導，データの取り方や分析の仕方まで丁寧に指導していただいた．一口に言えば，論文の作成全般（大学教員として自信となる何か）に対し，大変お世話になりました.

また，学会発表時や研究に対するヒントなど工学研究科の平田廣則，小圷成一両教授や日頃，輪講で討論いただいた研究室の方々，院生の方々に感謝いたします.

また，敬愛大学のメディアセンターの方々，特に，綱淵，五十嵐両氏には, MATLAB のプログラムインストールや図の変換，図の組み込みや PLATEX のインストール，使い方の助言など度々お世話になったことに感謝いたします.

最後に，敬愛大学国際学部の先生方に対しては，研究の意欲を与えてくれた先生，続けることに対して励ましの言葉をかけて下さった先生方に深く感謝いたします.