階層型ニューラルネットの学習特性に対する幾何学的解析に関する研究

白土　浩

目次

第1章　序論

人間の脳に近い知的情報処理の実現に向けて，ニューラルネットに関する研究が注目されている．ニューラルネットには，層から層への一方向性の結合を基本とした階層型ニューラルネットワーク（以降，多層ネットと略称）と，ユニット間で相互結合している相互結合型ネットワークに大別される．また，これらのネットワークの学習法は，誤差逆伝搬 (Back Propagation:BP) 法に代表される教師有り学習と自己組織化に代表される教師なし学習に分けられる．この他にもニューラルネットには種々のモデルや学習アルゴリズムがあり，これらの組み合わせによって，パターン認識や関数近似，最適化問題などの分野に応用されている．なかでも，多層ネットを BP 法で学習させる方法は，パターン認識や制御などの様々な分野で応用されており，その有用性は広く認められている．しかし，この最も単純で広く用いられている多層ネットと BP 法の組み合わせに対する収束性でさえ解明されていない．

多層ネットの学習に関する基本原理は，1957 年に Rosenblatt[28] によってパーセプトロンとして考案された．このモデルは，ユニット特性を閾値関数とした三層からなるネットに対して，出力ユニットの荷重を調整することで学習を行うものである．しかし，パーセプトロンでは中間層に対して適切な学習法が提示されていないことから，Minsky ら[24] は，XOR 問題などの線型分離不可能問題に対しては解を得ることができないとして，パーセプトロンの学習能力の限界を指摘している．この中間層の学習法については，甘利[31),44)] がユニット特性を微分可能なアナログ関数とし，最急降下法によりネット出力と教師との自乗誤差を最小化するように荷重を調整する学習法を提案している．その後，Rumelhart ら[1),5)] は，これとは独立に，ユニット特性を非線形関数（一般にはシグモイド関数）として同様の原理に基づく学習法を BP 法として提案している．そして，現在，BP 法は多層ネットの代表的な学習アルゴリズムとして用いられている．しかし，BP 法は，最急降下原理に基づいているため，必ずしも大域解 (Global Minima) に収束するとは限らず局所

最小値 (Local Minima) へ落ち込むことや学習速度が遅いことなどの収束上の問題がある．さらに，三層ネットは，無限個の中間ユニットを許せば任意の連続実関数を任意の精度で近似できることが保証されている[9),17)]が，各層のユニット数を定める方法が確立されていないため，その運用は試行錯誤にならざるを得ないといった欠点もある．このような収束問題は，現実的な課題を大規模のネットで学習させる場合，更に深刻さを増すものと考えられる．

収束問題については，BP 法は最急降下原理に基づく多峰的な誤差曲面の最小値探索に他ならないため，誤差曲面の最適化の観点から，収束速度の向上を目指した方法として，学習係数を最適化する方法[13),14),39)]や Quickprop 法[32)]，その他にも種々の高速化手法[2),27),41),42)]が提案されている．一方，同様の理由から，BP 法による学習が収束するか否かは初期値に大きく依存する．そのため，この観点から収束能力の改善を目指した初期値設定法として，主成分分析[16)]やクラスタ分析[47)]の観点に基づいて設定する方法や初期荷重の分布幅をユニットへの入力本数[30)]やユニット数[6)]により定める方法などがある．このように，アルゴリズムや初期値設定によって収束能力の改善が試みられているが，収束性の本質的な議論に関しては解明されていないのが現状である．

本来，収束性を厳密に議論をするには，学習アルゴリズムの挙動を記述する微分方程式に基づいて，解の安定性を位相空間図でのフローやアトラクタに基づいて解析する必要がある[21),35)]．このような解析手法は線形システムに対して有効であるが，多層ネットのように非線形システムの場合，2.3.1 節でも述べるように，解のアトラクタは学習課題に依存して一般には未知であるため，収束挙動を厳密に議論することは難しい．これは，多層ネットの収束挙動を解析した報告例が見当たらないことからも推察される．したがって，本論文では，収束性について言及する場合，時間軸に対する収束挙動そのものを含むものではなく，初期値を通常の一様乱数で与えた場合の学習試行における解への収束能力を指すものとする．

一方，多層ネットには訓練データに対する学習収束能力だけでなく，未学習データに対する識別能力である汎化能力も要請される．この汎化能力については，一般に中間ユニット数が必要以上に多くない方が良好であるとされている[12)]．最適な中間ユニット数の決定法としては，情報量基準 AIC(Akaike's Information Criterion)[11)]に基づいた方法[36),46)]や種々の発見的方法が提案されている[15),18)]．また，最適な中間ユニット数を学習によって定める手法としては，学習開始時に十分な数の中間ユニットを用意しておきユニットや荷重を学習過程で削除していく削除的学習法[29)]

と，小さなネットで学習を開始して学習過程においてユニットを順次増やしていく生成的学習法がある．削除的学習法には，荷重やユニットをある評価基準に基づいて削除していく方法[4),45),52)]と，付加的評価を用いる方法として，石川[25)]の忘却付き構造学習や安井[34)]の相互抑制型学習などがある．

以上のように，多層ネットの収束性の改善や汎化能力の向上に関する研究は多くなされているが，依然として収束性に関する本質的な問題については解明されていない点が多い．例えば，ユニット特性の比較として，シグモイド関数の活性範囲が 0 から 1 のように正のみの値をとる片極型のユニットからなる片極型ネットと，-1 から 1 のように正負の両方の値をとる両極型のユニットからなる両極型ネットとの比較で，結合荷重やバイアス荷重の初期値を通常のように平均ゼロの一様乱数で与えた場合，両極型による方が収束は速いとされている[8),12)]．また，三層ネットの中間層のみを両極型のユニットとすることによって，収束速度やパターンの分離能力が向上したり[26)]，解の符号の組み合わせが多くなって収束率が改善されるとの実験報告もある[33)]．このように実験的もしくは経験的に両極型ネットの収束上の優位性を指摘している報告例は多く存在するが，学習特性や収束性の観点から，この問題を体系的に議論している文献は見当たらない．

収束性に関しては，BP 法が最急降下原理に基づく多峰的な誤差曲面の最小値探索であることから，評価関数をパラメータ（荷重）空間上の誤差曲面とみなして議論を展開することが多い．しかし，XOR 問題を入力ユニットが 2 個，中間ユニットが 2 個，出力ユニットが 1 個の三層ネットで学習させるという簡単な場合でも，パラメータである荷重は 9 個必要となる．したがって，誤差曲面に対するパラメータ空間は 9 次元空間となり，誤差曲面の形状を知ることは困難となる．そのため，実際の解析例としては，XOR 問題を対象に，9 個のパラメータのうち 7 個を固定したネットに対する解のアトラクタを例示した報告[54)]や，荷重によって形成される記憶面と呼ばれる超平面の 2 次元断面図を用いて例示的に学習過程を解析した報告[37),53)]に留まっている．しかし，収束性を議論するためには，より一般的な問題を多層ネットで解く場合について検討する必要がある．この場合，ネット構造は更に複雑になり，学習に必要な荷重の本数，等価的に誤差曲面に対するパラメータ空間が増大するため，誤差曲面に基づいた議論は困難となる．

賈ら[40)]は，分離超平面と活性領域を用いて，ユニットの情報伝達構造を表現する特徴パラメータを導入し，片極型ネットの初期値を乱数で与えられるときの特徴パラメータの漸近特性を調べ

ることで，多層ネットの初期状態について考察を行っている．そして，ユニットへの入力本数が増えると，活性領域の幅が入力定義域のサイズと整合しなくなったり，分離超平面が原点に集中するため，荷重更新が円滑に進まなくなることを指摘している．さらに，この問題点を解消するため，各ユニットの分離超平面が入力定義域の中心を通るように初期バイアス荷重を定めて，その活性領域が入力定義域を適切に覆うように初期荷重を設定する方法を提案している．また，これを発展した設定法として，荷重を全て正の乱数で与えることで，さらに収束性が向上するとの報告[38]もある．

上述の幾何学的観点からの議論をさらに発展させることで多層ネットの学習収束を次のように解釈することが可能となる．すなわち，多層ネットによる学習が収束して解となる荷重が得られた場合，荷重の幾何学的表現である分離超平面と活性領域は提示入力パターンの幾何学的表現である入力集合を有意義に分離している必要があり，この条件は，多層ネットが解を持つための必要条件であると捉えることができる．したがって，解の必要条件が解析的に求まれば，必要条件と初期値を解のアトラクタの観点に基づいて解析することで静的な学習特性に関する議論が可能となる．また，解が得られた場合，分離超平面は本質的に入力集合の内部を通る必要があることから，解に対する探索領域をユークリッド空間全体から必要条件を満たす領域へと絞り込むことが可能となる．

しかし，入力集合は学習課題に依存しており，仮に入力集合が特定できても，その内部を有意義に分離する分離超平面を特定することは難しく，解の必要条件を厳密に導出することは困難である．そこで，必要条件に代わる近似的な条件として，分離超平面が入力集合の外接直方体の内部を通る条件が求まれば，収束性に関する厳密な議論が可能となると考えられる．しかし，近似条件を分離超平面で幾何学的に記述できたとしても，分離超平面が集合表現であるため定性的な範疇でしか議論できない．

本論文では，この幾何学的な条件をより定量的に議論するために分離超平面に随伴する法ベクトルの概念を新たに導入し，分離超平面が入力集合の外接直方体の内部を通るための条件を法ベクトルの集合により定式化する．そして，この解の必要条件の近似的な条件を分離条件と呼ぶことにする．また，分離条件や初期値を法ベクトルの分布として可視化することにより，必要条件の高次元ベクトル空間における幾何学的な形状に関する知見が得られる．この知見をもとに，入

力パターン集合が単に平行移動した場合やネット極性が違う場合でも，分離条件は大きく異なることを明らかにする．そして，分離条件と初期値を同一ベクトル空間上で比較することで，これまで定性的にしか示されていなかった両極型ネットの片極型ネットに対する収束上の優位性を裏付ける．さらに，両極型ネットと片極型ネットの許容解濃度に関して検討し，解濃度は極性に拠らず提示入力パターンがアフィン変換の関係にあれば同一となり，分離能力も等しくなることを明らかにする．この結果をもとに，多層ネットをハードウェア化する際の設計指針を提案する．さらに，法ベクトルの概念に基づいた三層ネットの初期値設定法についても考察を行う．

次章以降の内容について順を追って説明し，本論文の構成を明らかにする．

まず，第2章では，従来の誤差曲面からのアプローチに基づいた議論では，収束性に関して掘り下げた議論が困難となることを述べる．そこで，誤差曲面からのアプローチに代えて，幾何学的観点からのアプローチに基づいた収束性に関する議論を展開する．そして，原点から伸びて分離超平面と直交する法ベクトルの概念を新たに導入し，両極型ネットと片極型ネットでは，初期値が通常のように乱数で与えられる場合，分離超平面は極性に拠らず同一となるが，入力定義域は極性により幾何学的に異なるため，分離超平面による入力定義域の分割が異なることを明らかにする．

そして，第3章では，学習開始時における入力定義域分割の極性による違いが，荷重更新や解の実現にどのように影響するかを調べて，片極型ネットの場合，両極型ネットに比べて，良好な収束を与える初期値の範囲が狭くなる傾向にあることや，得られる解の種類が偏る恐れがあることを指摘する．さらに，両極型ネットと片極型ネットでXOR問題を学習させる場合の解領域を図示し，極性によって解領域は大きく異なることを明らかにする．そして，ネットサイズが大きな場合，片極型ネットに比べて両極型ネットの方が，より広い範囲の初期値に対して良好な収束を与えることをシミュレーションにより確認する．さらに，片極型ネットと両極型ネットで，分離超平面による入力定義域の分割を幾何学的に等価にした場合，提示パターンに対する誤差特性は同じになることを指摘する．これは，解領域の大きさがネットの極性によらず同一であることを意味している．そして，学習開始時での入力定義域の分割が幾何学的に等価でも，荷重の更新特性はユニットの極性によって異なるため，得られる収束解や収束に至るまでの学習回数などの収束能力は，片極型ネットと両極型ネットとで異なることを述べる．そして，シミュレーション

結果をもとに，入力定義域の初期分割が等価でも，ネットサイズが大きな場合，両極型ネットは，片極型ネットに比べて，良好な収束を示す傾向にあることを明らかにする．

つぎに，第4章では，第3章で述べた解領域の大きさに関してより厳密に検討する．そして，入力パターンがアフィン変換の関係にある学習課題に対して，層数や対応する層のユニット数が同じ多層ネットは，シグモイド関数の極性や値域幅の大きさによらず，等しい濃度の許容解をもつことを示す．この結果を利用して，両極型ネットを対象に，入力パターンやシグモイド関数の値域が k 倍された場合，つまり入力定義域が k 倍に拡大された場合，バイアス荷重を同じにして，荷重を $1/k$ 倍すれば，分離能力や汎化能力を等価に保存できることを指摘する．さらに，荷重の初期値を上記のように設定した場合，バイアス荷重の更新に対する学習係数を $1/k^2$ 倍，また荷重更新に対する学習係数を $1/k^4$ 倍にすれば，収束挙動は等価になることを述べる．最後に，通常のように，荷重の初期値を同一分布幅の一様乱数で設定し，双方の更新に対する学習係数を等しく与えて学習させた場合，良好な収束を与える初期値は入力定義域の拡大とともに小さくなることをシミュレーションにより示す．

第5章では，前章までの議論を踏まえて，多層ネットの収束解と初期値を同一ベクトル空間上で議論することにより，両極型ネットの片極型ネットに対する収束上の優位性を明らかにする．その際，幾何学的観点から見た多層ネットの学習収束である分離超平面が入力集合を有意義に分離する条件が解析的に求まれば，収束性に関して厳密な議論が可能となるが，そのような条件を求めることは困難である．そこで，入力集合の代わりにこれに外接する（超）直方体で近似し，その内部を分離超平面が通るための条件を分離条件と呼ぶことにして，これを法ベクトルの集合により定式化する．そして，この分離条件を満たす法ベクトルの分布領域を2次元の場合について図示する．この図から，高次元ベクトル空間における分離条件の幾何学的な形状に関する知見が得られ，入力パターン集合が単に平行移動したり，ネット極性が違うだけでも，分離条件は大きく異なることがわかる．また，荷重の初期値を通常のように平均がゼロの一様乱数で与えたときの法ベクトルの分布領域についても図示する．以上の二つの分布領域を同一のベクトル空間で比較することにより，通常の初期値設定によるBP学習の収束は，片極型ネットに比べて，両極型ネットの方が優位となることが導かれる．さらに，双方のネットで入力集合の分割が等価となるように初期値設定した場合でも，片極型ネットに比べて，両極型ネットによる収束は幅広い学習

係数に対して良好となることをシミュレーションにより示す．最後に，分離条件を考慮した初期値設定の効果について考察する．

第6章では，パターン分類問題を三層両極型BPネットで学習させる際の初期値を幾何学的観点に基づいて設定する方法について考察する．すなわち，まず，中間層と出力層における入力集合の性質が異なるため，両者に対する初期値設定を分けて行う必要があることを述べる．次に，中間層については，そのユニットのなす分離超平面が提示入力集合を内部を通るように荷重を設定し，出力層については，先験情報が偏って与えられないように，すべての荷重をゼロと設定する．パリティ問題やMONK's問題，ソナー問題，アヤメの分類問題に対するシミュレーションの結果，中間層に対する初期分離超平面が入力集合の中心を通る場合に最も良好な結果が得られることや提案法を用いることで通常のように平均ゼロの一様乱数で初期値設定した場合に比べて，収束性が改善されることを明らかにする．

以上について，第7章では総括する．

第2章　幾何学的アプローチによる収束性解析

2.1　緒言

多層ネットによるBP学習は，パターン認識や制御等の多くの分野に応用されており，その有用性は広く認められている．しかし，BP法は，最急降下原理に基づいているため，必ずしも求める解(Global Minimum)に収束するとは限らず局所最小値(Local Minima)への落ち込みや，極小値付近での荷重更新量が小さくなるため学習速度が遅い，といった収束上の問題がある．さらに，適切なネットサイズが未知であるため，その運用は試行錯誤にならざるを得ないといった欠点もある[12]．そして，このような収束問題は，現実的な課題を大規模のネットで学習させる場合，深刻さを増すと考えられる．

収束問題については，BP法のアルゴリズムや，初期値設定の観点から様々な検討がなされている．まず，BP法は最急降下原理に基づく多峰的な誤差曲面の最小値探索に他ならない．したがって，誤差曲面の最適化の観点から，収束速度の向上を目指した方法として，学習係数を最適化する方法[13),14),39)]やQuickprop法[32)]，その他にも種々の高速化手法[2),27),41),42)]が提案されている．一方，多層ネットには，学習データの規定する入出力関係を満たす解が多数存在すため，誤差曲面から見て解のアトラクタ領域に高い確率で初期荷重を設定できれば，学習の収束性が向上すると考えられる．この観点から，初期値設定の学習収束に及ぼす影響に関する研究もなされている[20)]．そして，初期値設定法には，荷重を訓練データに対する主成分分析[16)]やクラスタ分析[47)]の結果に基づいて設定する方法や，初期荷重の分布幅をユニットへの入力本数[30)]やユニット数[6)]により定める方法などがある．このように，アルゴリズムや初期値設定によって収束性が改善されてるが，その本質である収束性に関しては解明されていないのが現状である．

本章では，本論文での基本原理となる分離超平面に随伴する法ベクトルを定義し，幾何学的ア

プローチに基づいた収束性に関する議論についての基本概念を述べ，以降の章に対する準備のための章とする．まず，両極型ネットと片極型ネットについて定義を行い，従来の誤差曲面からのアプローチによる議論では，収束性に関して掘り下げた議論が困難となることを述べる．そこで，誤差曲面からのアプローチに代えて，幾何学的観点からのアプローチにより収束性を議論する．すなわち，多層ネットによる学習が収束して解となるためには，荷重の幾何学的表現である分離超平面と活性領域が提示入力パターンの幾何学的表現である入力集合を有意義に分離する必要がある．そのため，少なくとも分離超平面は入力が存在し得る範囲（入力定義域）の内部を通る必要がある．したがって，従来の誤差曲面からのアプローチでは，収束解をパラメータ空間全体に渡って探索する必要があるのに対して，幾何学的アプローチでは，解の探索空間をユークリッド空間全体から上述の条件を満たす領域へと絞り込むことが可能となる．そして，両極型ネットと片極型ネットでは，初期値が通常のように乱数で与えられる場合，分離超平面は極性に拠らず同一となるが，入力定義域は極性により幾何学的に異なるため，分離超平面による入力定義域の分割が異なることを明らかにする．

2.2 両極型ネットと片極型ネット

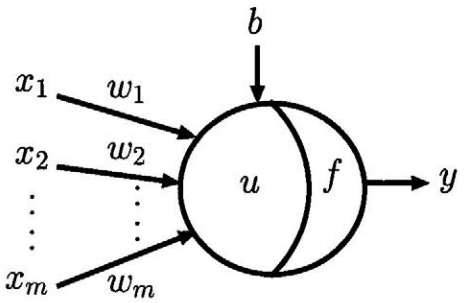


図 2.1: ユニットの入出力特性

多層ネットのあるユニットの入出力特性を

$$u = \sum_{i=1}^{m} w_i x_i + b \tag{2.1}$$

$$y = f(u) \tag{2.2}$$

のように定義する．ここに，b はバイアス荷重，w_i は一つ下位の層の第 $i(=1,2,\cdots,m)$ ユニットとの結合荷重，x_i はそのユニットからの入力である (図 2.1)．そして，シグモイド関数 $f(\cdot)$ が

$$f_U(u) = \frac{1}{1+e^{-u}} \tag{2.3}$$

$$f_B(u) = \frac{1}{1+e^{-u}} - \frac{1}{2} = \frac{1}{2}\tanh(\frac{u}{2}) \tag{2.4}$$

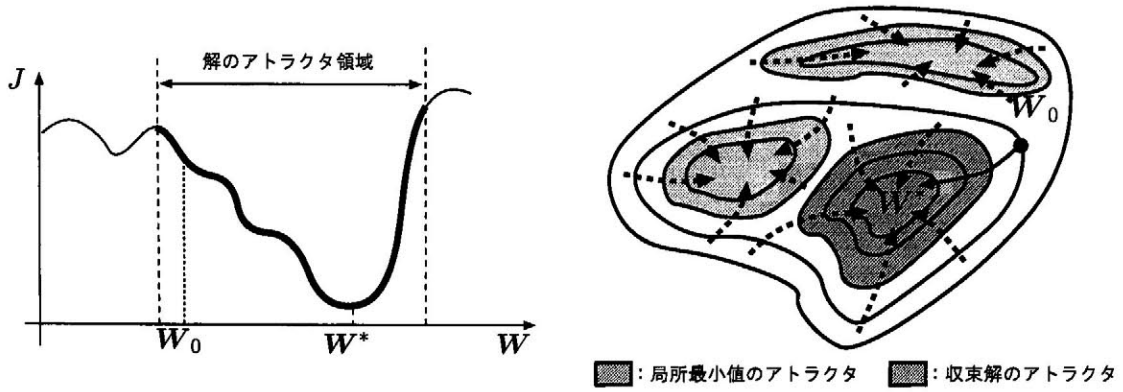
で定義されるユニットをそれぞれ片極型ユニットと両極型ユニットと区別する．そして，前者により構成される多層ネットを片極型ネット，後者により構成される多層ネットを両極型ネットと呼ぶ．このとき，変数や集合は必要に応じて，(2.3)(2.4) 式のように，両極型ネットに関するときを添字の B で，片極型ネットに関するときを添字の U で区別する．さらに，両者は同一構造，すなわち，対応する層のユニットは同数とする．

2.3　収束性の解析手段

ここでは，従来の誤差曲面からのアプローチによる議論では，収束性に関して掘り下げた議論が困難となることを指摘する．そして，誤差曲面からのアプローチに代えて，幾何学的観点からのアプローチにより収束性を議論する．すなわち，幾何学的観点からみた学習収束を，分離超平面が訓練データやユニット出力からなる入力集合を有意義に分割することであると捉え，これを解の必要条件とみなすことで，収束性について議論する．

2.3.1　誤差曲面からのアプローチ

多層ネットの収束問題に関する研究は多くなされているが，その本質である収束性に関しては未だに解明されていないのが現状である．収束性に関しては，BP 法が最急降下原理に基づく多峰的な誤差曲面の最小値探索であることから，評価関数をパラメータ（荷重）空間の誤差曲面とみなして議論を展開することが多い．以降では，このような議論を誤差曲面によるアプローチと呼ぶことにする．



(a) 誤差曲面の模式図　　(b) (a) の等高線表示とパラメータ空間

図 2.2: 誤差曲面から見た初期値と収束解およびアトラクタ領域

図 2.2 は誤差曲面から見た初期値と収束解および収束解のアトラクタ領域を模式的に表した図である．2.2(a) の横軸は多層ネットの全荷重からなるパラメータ空間 $\boldsymbol{W}$，縦軸はそれに対応する評価関数 J を表す．また図 2.2(b) は，$\boldsymbol{W}$ を 2 次元として，2.2(a) の誤差曲面を等高線表示したものに初期値から解への収束挙動を示すフローと解のアトラクタを模式的に描いた位相空間図を表している．図から，誤差曲面から見た学習収束とは，多層ネットの初期荷重 $\boldsymbol{W}_0$ が最急降下原理により収束解である $\boldsymbol{W}^*$ へ収束していくことである．すなわち，多層ネットの学習が収束するか否かは，$\boldsymbol{W}_0$ が解のアトラクタに位置するか否かによって決定される．したがって，誤差曲面における解のアトラクタが解析的に求まれば収束挙動を含む収束特性を厳密に議論できる．しかし，解のアトラクタ領域は学習課題に依存しており，仮に学習課題を固定した場合でも多層ネットが非線形構造を有しているためアトラクタを特定することは難しい．したがって，図 2.2(b) における解のアトラクタやフローに関する知見が得られないため位相面解析や学習挙動を記述した微分方程式における解の安定性に基づいた時間軸を含む収束挙動解析は困難となる．

一方，収束挙動を含まない収束性に関する議論についても，XOR 問題を入力ユニットが 2 個，中間ユニットが 2 個，出力ユニットが 1 個の三層ネットで学習させるという簡単な場合でも，パラメータである荷重は 9 個必要となる．したがって，図 2.2 の誤差曲面におけるパラメータ空間 ($\boldsymbol{W}$) は 9 次元空間となり，誤差曲面の形状を知ることは困難となる．そのため，実際の解析例と

しては，XOR問題を対象として，9個のパラメータのうち2個を可変とし，残りを固定したネットに対して解のアトラクタを例示した報告[54]や，荷重によって形成される記憶面と呼ばれる超平面の2次元断面図を用いて例示的に学習過程を解析した報告[37),53]に留まっている．しかし，より一般的な問題を多層ネットで解く場合，ネット構造は更に複雑になり，学習に必要な荷重が増大するため，誤差曲面に準じた議論は困難となる．これは，従来の誤差曲面からのアプローチでは，収束解をパラメータ空間全体に渡って探索する必要があることに起因していると考えられる．したがって，XOR問題を解く場合でも上述のことから探索空間は広大なものとなり，ネット構造が複雑になると共にパラメータ空間も増大し，探索空間に対する知見，等価的に誤差曲面の形状を知ることが困難となることを示唆している．

2.3.2　幾何学的観点からのアプローチ

前節で述べたように誤差曲面に基づいた議論では，本質的な収束性に関して掘り下げた議論は困難となる．賈ら[40]は，片極型ネットを対象として，ユニットの情報伝達構造を表現する特徴として，分離超平面と活性領域を導入した．

$$L_p(0.5) : w_0 \quad + \quad \sum_{i=1}^{m} w_i x_i = f_U^{-1}(0.5) \tag{2.5}$$

$$\{(x_1,\ x_2,\ \cdots,\ x_m) \quad ; \quad \rho < f\left(\sum_{i=1}^{m} w_i x_i + w_0\right) < 1-\rho\} \tag{2.6}$$

ここに，ρ は活性領域を規定する 0.1 程度の小さな値である．そして，ユニット特性が片極型で，初期値が乱数で与えられるときの特徴パラメータの漸近特性を調べることで，多層ネットの初期状態について考察を行っている．その結果，ユニットへの入力本数が増えると，活性領域の幅が入力定義域のサイズと整合しなくなったり，分離超平面が原点に集中するため，荷重更新が円滑に進まなくなることを指摘した．

上述の分離超平面や活性領域の極性を考慮したものを，それぞれ H_U, H_B, $\mathcal{G}_U$, $\mathcal{G}_B$ とすれば，

$$H_B = H_U \quad = \quad \{\boldsymbol{x} \mid \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b = 0\} \tag{2.7}$$

$$\mathcal{G}_U \quad = \quad \{\boldsymbol{x} \mid \rho < f_U(u) < 1-\rho\} \tag{2.8}$$

$$\mathcal{G}_B \quad = \quad \{\boldsymbol{x} \mid -0.5+\rho < f_B(u) < 0.5-\rho\} \tag{2.9}$$

となって，分離超平面は極性に拠らず同一の表現となり，活性領域については両者で表記は異なるものの R^m に占める割合は等しくなることが分かる[49]．ここに，$\boldsymbol{x} = [x_1, x_2, \cdots, x_m]^{\mathrm{T}}$, $\boldsymbol{w} = [w_1, w_2, \cdots, w_m]^{\mathrm{T}}$，T は転置記号を表す．

賈らの議論をさらに発展させることで，多層ネットの学習収束を幾何学的観点からのアプローチとして，以下のように捉えることができる．図 2.3 はユニットの入力を 2 次元としたときの分離超平面による入力集合の分割を表している．図中，横軸と縦軸はユニットへの入力 x_1, x_2 を表している．この図から，多層ネットによる学習が収束して解（収束解）となるためには，初期荷重 $\boldsymbol{w}_0$ の幾何学的表現である分離超平面 H_0 が，2 つのクラス C_1, C_2 からなる入力 $\boldsymbol{x}$ の幾何学的表現である入力集合 $\mathcal{X}$ を有意義に分離する分離超平面 H^* へ推移していくことであると解釈できる．したがって，少なくとも収束解となる分離超平面 H^* は入力が存在し得る範囲（入力定義域 X）の内部を通る必要があるため，収束解の探索空間をユークリッド空間 (R^m) 全体から上述の条件を満たす領域へと絞り込むことが可能となる．

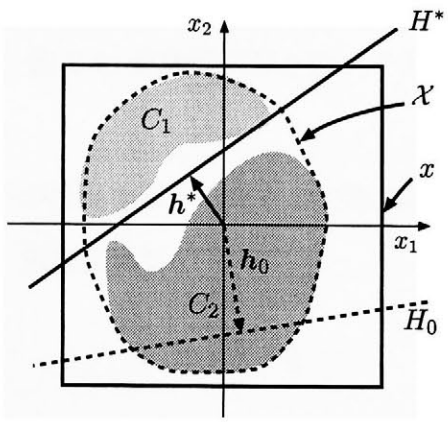


図 2.3: 幾何学的観点からみた学習収束

しかし，分離超平面や活性領域は集合表現であるため，これらを用いた議論では，定性的にしか収束性を議論できない．そこで，次節では分離超平面に随伴する法ベクトル $\boldsymbol{h}$ を導入する．

2.4　分離超平面から法ベクトルへ

ここでは，分離超平面に随伴する法ベクトルを以下のように求める．図 2.4 は両極型のときの

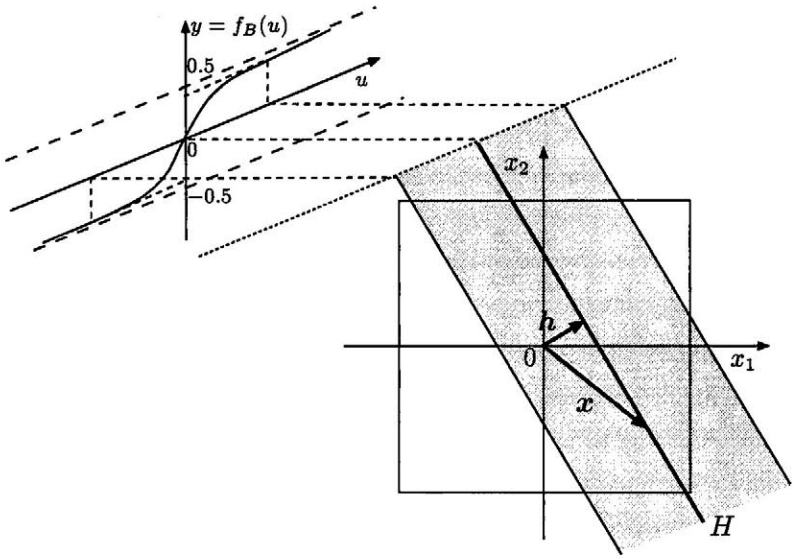


図 2.4: 分離超平面と法ベクトル

シグモイド関数と入力定義域との関係を図示したものである．分離超平面は総入力がゼロとなる点の集合として表されるため，極性に拠らず

$$\boldsymbol{w}^T\boldsymbol{x} + b = 0 \tag{2.10}$$

を満たす入力 $\boldsymbol{x}$ の集合として表される．

いま，原点から伸びて分離超平面と直交するベクトルを法ベクトル $\boldsymbol{h}$ とすれば，$\boldsymbol{h}$ は

$$\boldsymbol{w}^T\boldsymbol{h} + b = 0 \tag{2.11}$$

$$\boldsymbol{h}^T(\boldsymbol{x} - \boldsymbol{h}) = 0 \tag{2.12}$$

なる関係を満たすことになる．そこで，(2.10) 式から (2.11) 式を引くと

$$\boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{h}) = 0 \tag{2.13}$$

となり，これと (2.12) 式を比較することで，

$$\boldsymbol{h} = \beta \boldsymbol{w} \tag{2.14}$$

となることが分かる．そして，(2.14) 式を (2.11) 式に代入することで

$$\boldsymbol{w}^T(\beta \boldsymbol{w}) + b = 0 \tag{2.15}$$

となって，β は

$$\beta = -\frac{b}{\|\boldsymbol{w}\|^2} \tag{2.16}$$

のように求まる．したがって，法ベクトル $\boldsymbol{h}$ は

$$\boldsymbol{h} = -\frac{b}{\|\boldsymbol{w}\|^2}\boldsymbol{w} \tag{2.17}$$

と記述される．また，活性領域の幅[40]についても極性に拠らず，

$$g = \frac{2\ln\{(1-\rho)/\rho\}}{\|\boldsymbol{w}\|} \tag{2.18}$$

となる．

以降では，幾何学的な表現としての，分離超平面 ((2.7) 式) や活性領域 ((2.8)(2.9) 式) の代わりに，法ベクトル ((2.17) 式) や活性領域幅 ((2.18) 式) に基づいて収束性に関する議論を進めていく．

2.5 極性による入力定義域分割の相違

分離超平面による R^m の分割については，極性による差異はない．しかし，入力定義域の分割について見てみると，極性の違いにより以下のような差異が生じる．すなわち，入力定義域は，片極型のとき $X_U = (0,1)^m$，両極型のとき $X_B = (-0.5, 0.5)^m$ で与えられ，いずれも同じ大きさの単位超立方体となって，その対角長が

$$D = \sqrt{m} \tag{2.19}$$

と定義される入力定義域サイズも極性によらず等しい．しかし，X_U が R^m の原点 $\boldsymbol{o}$ を頂点とする第 1 象限内の単位超立方体となるのに対して，X_B は R^m の原点を中心とする単位超立方体

となって，R^m に対する入力定義域の位置付けは極性によって異なる．一方，荷重の値を片極型と両極型で等しく与えた場合，(2.17) 式で定義される $\boldsymbol{h}_j$ は極性によらず R^m で同じベクトルとなって同一の分離超平面 H_j を指すことになる．そのため，分離超平面による入力定義域の分割

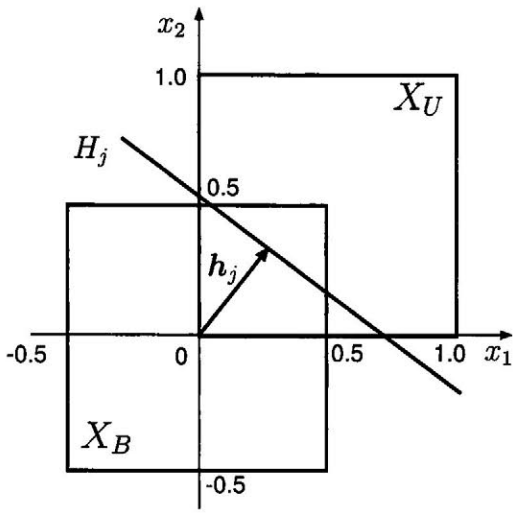


図 2.5: 分離超平面 (H_j) による入力定義域 $(X_B,\ X_U)$ の分割

は，図 2.5 に示されるように，両者には差異が生じる．すなわち，図 2.5 において，分離超平面 H_j は，両極型の入力定義域 X_B の右上角を通るのに対して，片極型の入力定義域 X_U の左下角を通る，という違いがある．

2.6　結言

本章では，本論文での基本原理となる分離超平面に随伴する法ベクトルを定義し，幾何学的アプローチに基づいた収束性に関する議論についての基本概念を述べ，以降の章に対する準備を行った．すなわち，まず，両極型ネットと片極型ネットについて定義を行い，従来の誤差曲面からのアプローチによる議論では，一般的な問題を対象とした収束性に関する議論が困難となることを指摘した．そして，誤差曲面からのアプローチに代えて，幾何学的観点から学習収束を，分離超平面が訓練データやユニット出力からなる入力集合を有意義に分割することであると捉え，これ

を多層ネットが解を持つための必要条件と考えることで，収束性に関する議論が可能となることを述べた．しかし，分離超平面や活性領域は集合表現であるため，これらに基づいたままでは定性的にしか収束性に関する議論を展開できない．そこで，分離超平面に随伴する法ベクトルの概念を新たに導入した．そして，両極型ネットと片極型ネットでは，初期値が通常のように乱数で与えられる場合，分離超平面は極性に拠らず同一となるが，入力定義域は極性により幾何学的に異なるため，分離超平面による入力定義域の分割が異なることを明らかにした．

以降の章では，幾何学的観点からのアプローチにより，初期値や収束解を (2.17) 式で表される法ベクトルや (2.18) 式で表される活性領域幅に基づいて収束性に関する議論を進めていく．

第3章　ネット極性と収束挙動

3.1　緒言

誤差逆伝搬 (BP) 学習法は，パターン認識や制御等の多くの分野でその有効性が報告されているが，学習速度の遅さや局所最小値への落ち込みなどの問題があることも良く知られている．そのため，種々の観点から収束の改善が試みられており，その１つに，ユニット特性の比較からの報告がある．すなわち，シグモイド関数の活性範囲が 0 から 1 のように正のみの値をとる片極型のユニットからなるネットワークと，−1 から 1 のように正負の両方の値をとる両極型のユニットからなるネットワークとの比較で，荷重やバイアスの初期値を乱数で与えた場合，両極型による方が収束は速いとされている[8),12)]．また，３層ネットワークの中間層のみを両極型のユニットとすることによって，収束速度やパターンの分離能力が向上したり[26)]，解の符号の組み合わせが多くなって収束率が改善されるとの実験報告もある[33)]．しかし，このように両極型による収束が優位となる根拠は明確でない．

本章では，前章で述べた学習開始時における入力定義域分割の極性による違いが，荷重更新や解の実現にどのように影響するかを 2 − 2 − 1 ネットで XOR 問題を解かせる場合について調べ，片極型の場合，両極型に比べて良好な収束を与える初期値の範囲が狭くなる傾向にあることや，得られる解の種類が偏る恐れがあることを指摘する．また，ユニット特性がしきい値関数で与えられる場合について，XOR 問題の解領域を法ベクトルの集合で定式化して階層間の関係を明らかにする．そして，解領域がユニット極性によって大きく異なることを指摘する．この結果から，ユニット特性がシグモイド関数で与えられる通常の多層ネットで一般的な課題を BP 学習させる場合，アトラクタや解領域はユニット極性に依存することが示唆される．つぎに，ネットワークサイズが大きい場合，両極型による方が，より広い範囲の初期値に対して良好な収束を与えるこ

とをシミュレーションにより確認する．さらに，片極型と両極型ネットワークで，分離超平面による入力定義域の分割を幾何学的に等価にした場合，提示パターンに対する誤差特性は同じになることを指摘する．そして，学習開始時における入力定義域の分割が幾何学的に等価でも，荷重の更新特性はユニットの極性によって異なるため，得られる収束解や収束に至るまでの学習回数などの収束能力は，片極型ネットと両極型ネットとで異なることを述べる．また，中間ユニットの極性が同じで，最終の出力ユニットのみの極性が異なる 2 つのネットを比較して，初期値が等しい場合，両者の収束挙動は一致することを導く．シミュレーションの結果，入力定義域の初期分割が等価でも，ネットサイズが大きい場合，両極型ネットは，片極型ネットに比べて，良好な収束を示す傾向にあることを明らかにする．

3.2 XOR 問題での分割と出現頻度の解析

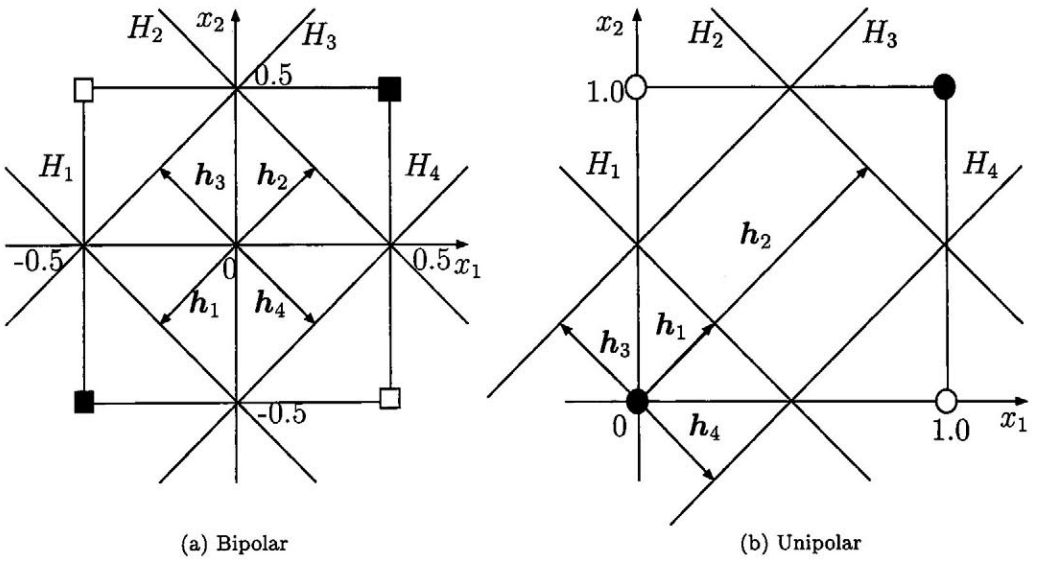


図 3.1: 幾何学的観点からみた XOR 問題の 2 種類の収束解

ここでは，m が小さくて，解となる分割の極性による違いが明らかな XOR 問題を $2-2-1$ のネットワークで学習させた場合を例に，分割と出現頻度の関係について大まかな様子を見るこ

とにする．

この場合，中間ユニットのなす分離超平面の対として，両極型のとき図3.1(a)のようにType A（H_1 と H_2 の対または $\boldsymbol{h}_1$ と $\boldsymbol{h}_2$ の対）と Type B（H_3 と H_4 の対または $\boldsymbol{h}_3$ と $\boldsymbol{h}_4$ の対）の2種類の解が存在し，片極型の場合にも，これと分割が幾何学的に等価な2種類の解が図3.1(b)のように存在する．

すなわち，両極型のときの入力定義域分割 (図3.1(a)) と，これと等価な片極型のときの入力定義域分割 (図3.1(b)) を見てみると，個々の分離超平面に対する比較から，H_3 や H_4 を指す $\boldsymbol{h}_3$ や $\boldsymbol{h}_4$ はそれぞれ極性によらず同じベクトルで記述されるものの，H_1 を指す $\boldsymbol{h}_1$ は極性によって方向が反対のベクトル，また H_2 を指す $\boldsymbol{h}_2$ は極性によって長さの異なるベクトルとして記述されることが分かる．

さらに，これらを全体的に比較して分かるように，両極型のときの $\boldsymbol{h}_j$ については，第1から第4象限までのすべての方向がとれて，すべて長さが入力定義域サイズの半分 $\sqrt{2}/2$ より小さくてすむのに対して，片極型のときの $\boldsymbol{h}_j$ については，第1，2，4の象限の方向に制限されて第3象限の方向をとることができない上に，$\boldsymbol{h}_2$ のように長さが $\sqrt{2}/2$ より大きいものも含まれる，と云った違いがある．また，片極型の $\boldsymbol{h}_j$ については，それが第2（第4）象限にある場合でも，図3.1(b)において x_1 軸からみた角度を θ とするとき，θ が $\pi(3\pi/2)$ に近づくにつれて，その長さ $\|\boldsymbol{h}_j\|$ は小さくなってゼロに漸近することから，その方向と長さは相互に強く依存して決まることが分かる．したがって，片極型の $\boldsymbol{h}_j$ については，それが第1象限以外に存在する場合，第1象限に存在するときに比べて，その存在範囲はより拘束されたものになる．他方，両極型の $\boldsymbol{h}_j$ については，それがどの象限にあろうと，その方向と長さの間の拘束関係は，第1象限に存在するときと同じである．

以上のことから分かるように，入力定義域を通る分離超平面に対する $\boldsymbol{h}_j$ は，両極型の場合，各象限に同じ割合で存在するが，片極型の場合，大部分が第1象限に偏在することになって，その存在範囲は両極型に比べて制約されることになる．

いま，初期荷重が平均ゼロの乱数で与えたときの初期分割を考え，法ベクトルの初期値 $\boldsymbol{h}_j^0$ の R^2 での分布はどの象限でも等しいと仮定して，Type A や Type B の分割との類似度を見てみる．まず，両極型の場合，図3.1(a)のように，$\boldsymbol{h}_3$ と $\boldsymbol{h}_4$ は，それぞれ $\boldsymbol{h}_1$ と $\boldsymbol{h}_2$ を反時計方向に

等しく $\pi/2$ だけ回転させたベクトルで，$\boldsymbol{h}_1$ と $\boldsymbol{h}_2$ の間の幾何学的な関係をそのまま保存している．したがって，両極型の場合，Type A と Type B に対して，類似度の同じ初期分割が等しい確率で存在することになる．一方，片極型の場合，図3.1(b) のように，$\boldsymbol{h}_3$ と $\boldsymbol{h}_4$ には，$\boldsymbol{h}_1$ と $\boldsymbol{h}_2$ の間の幾何学的な関係が保存されていないから，初期分割に対する Type A と Type B の類似度は異なる．また，Type B の $\boldsymbol{h}_3$ と $\boldsymbol{h}_4$ は，両極型と片極型で全く同じベクトルとなることから，Type B と初期分割の類似度は極性に依らず等しい．

学習はつぎの条件のもとで行った．すなわち，初期値は，幅が $(-\gamma, \gamma)$ の一様乱数を片極型と両極型で等しく設定し，訓練パターンについては，ユニットが片極型の場合，入力 x_i を 0 か 1 として教師を 0.1 か 0.9 で与え，両極型の場合，x_i を ± 0.5 として教師を ± 0.4 で与えた．また，慣性係数を $\alpha = 0.9$，繰返しの最大回数を 10000 回とし学習させて，全訓練パターンに対して，誤差の絶対値が 0.1 以内となったときを収束とした．そして，各 γ について，$N = 1000$ 回の乱数試行を行って，Type A (B) への収束数 N_A (N_B) を出現頻度として求めた．さらに，収束率 ($(N_A + N_B)/N$) と平均学習回数 (全収束数 $(N_A + N_B)$ における平均繰返し回数を訓練パターン数 4 で割った値) を求めた．表3.1に学習係数を $\eta = 0.6$ としたときの結果を示す．表3.1の g は学習開始時の活性領域幅 g_j^0，h は $\|\boldsymbol{h}_j^0\|$ の平均，D は入力定義域サイズを表し，ここでの値は $h = 0.875$，$D = \sqrt{2}$ である．

表3.1の結果より，両極型の場合，開始時の活性領域幅の平均 g が入力定義域サイズ D の 0.9 倍から約11倍の範囲で，Type A と B はほぼ等頻度で出現し，それ以上では，両方とも出現しないことが分かる．表3.1の結果より，両極型の場合，g が D の約11倍から 0.9 倍の範囲で，Type A と B はほぼ等頻度で出現していることが分かる．この結果は，双方の分割の初期分割からみた類似度が確率的に等価であることを反映している．一方，片極型の場合，g が D の 2 倍より小さい範囲では，Type B の方が先に見つかり易く，それより大きくなると，Type A の方が先に見つかり易くなって，両者の初期分割に対する類似度の違いが影響していることが分かる．つまり，2次元空間 R^2 が g の小さい明瞭な境界で分けられる場合と，g の大きい曖昧な境界で分けられる場合とで，Type A と B の実現の容易さは異なる．また，両極型と片極型で，初期分割に対する類似度が等しい Type B は，g が D の 3 倍より小さいときほぼ同じ頻度で出現するが，それより大きく11倍程度の範囲までは片極型の方の出現頻度が小さく，さらに50倍以上になると，

表 3.1: XOR 問題に対する両極型と片極型による学習結果
( ) 内は片極型による結果を表す. ($\eta = 0.6, h = 0.875, N = 1,000$)

初期値		収束率		平均学習回数		出現頻度	
γ	g/D	$(N_A + N_B)/N$				N_A/N_B	
0.01	542.5	0%	(4%)	–	(1606.7)	0/0	(41/0)
0.1	54.2	0%	(85%)	–	(371.0)	0/0	(850/0)
0.5	10.8	55%	(83%)	341.7	(244.7)	283/271	(658/176)
1.0	5.42	89%	(83%)	155.2	(168.5)	462/430	(465/374)
2.0	2.71	77%	(72%)	109.2	(151.2)	406/372	(333/387)
3.0	1.80	70%	(62%)	150.0	(194.5)	371/333	(271/355)
4.0	1.35	62%	(55%)	169.0	(266.7)	325/298	(238/318)
5.0	1.08	57%	(47%)	213.5	(248.0)	298/278	(204/267)
6.0	0.90	52%	(43%)	279.5	(336.7)	272/249	(183/252)

注; g/D は活性領域幅と入力定義域サイズの比で，N_A と N_B は解 Type A と B への収束数である.

両極型と片極型の双方で全く出現しなくなっている.

以上のことは，多種の解が存在するサイズの大きなネットワークの場合，片極型によって得られる解の種類は，両極型に比べて偏る可能性が高いことを示唆している．また，以上の結果から，類似度の評価において，$r(\boldsymbol{h}_j^0, \boldsymbol{h}_j^*)$ や $d(\boldsymbol{h}_j^0, \boldsymbol{h}_j^*)$ に対する重み付けは，g_j^0 が狭いとき大きく，逆に広いとき小さくなると考えられ，$q(g_j^0, g_j^*)$ に従属すると考えられる.

上述のように，Type A と B の出現頻度は極性によって異なるから，結果として得られる収束率や平均学習回数も極性によって異なることになる．すなわち，表 3.1 で良好な収束の得られる初期値の範囲をみてみると，片極型のとき $0.1 \leq \gamma \leq 2$ $(54.2 \geq g/D \geq 2.71)$，両極型のとき $1 \leq \gamma \leq 3$ $(5.42 \geq g/D \geq 1.80)$ で 7 割以上の収束率となっており，平均学習回数については，$\gamma = 0.5$ $(g/D = 10.8)$ で片極型の方が良好となるものの，$\gamma \geq 1$ $(g/D \leq 5.42)$ では両極型の方が良好となっている．他にも，次数が $m = 2, 3$ の低次のパリティ問題を解の存在が保証される最小のネットワーク $(m - m - 1)$ で $\eta = 0.2, 0.6, 1.0$ として学習させた結果，良好な収束率や平均学習回数を与える初期値の範囲は，上記と同じように片極型と両極型とで異なることが確認された．さらに，片極型ネットワーク $(2 - 2 - 1)$ の中間層のみを活性範囲が $(-1, 1)$ や $(-0.5, 0.5)$ の両極型ユニット[33),40)] に置き換えたネットワークについて，XOR 問題を γ や η を種々変えて学

習させた結果，これらの場合も良好な収束を与える初期値の範囲は，片極型ネットワークのときと異なることが分かった．

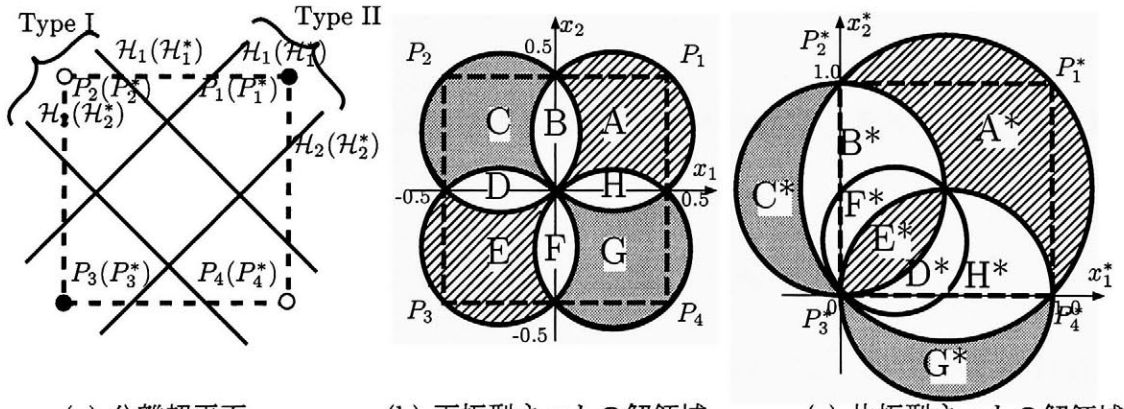
3.3 XOR 問題の解領域

ここでは，前節で述べた解の種類とその出現頻度について，より厳密な議論をするために，ユニット特性が閾値関数で両極型，片極型それぞれ

$$y = f(u) = \begin{cases} 1/2 & (u > 0) \\ -1/2 & (u < 0) \end{cases} \tag{3.1}$$

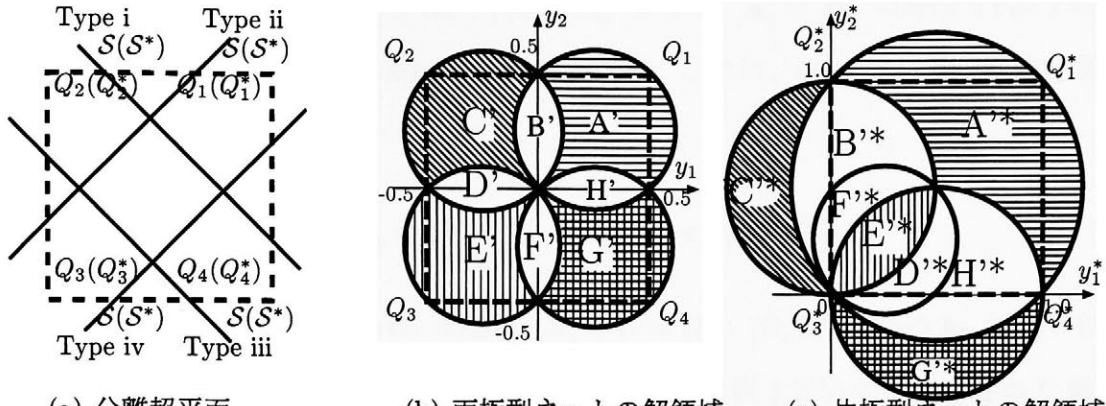
$$y^* = f^*(u) = \begin{cases} 1 & (u > 0) \\ 0 & (u < 0) \end{cases} \tag{3.2}$$

と与えられる 2 − 2 − 1 ネットによる XOR 問題の解について考察し，解領域がネット極性によって異なることを述べる．その際，議論の便宜を図るため，法ベクトルや分離超平面を次のように表記する．すなわち，(2.17) 式は，荷重やバイアスが双方で同じなら，ユニット極性に関係なく成り立つが，両極型と片極型ネットの中間層ユニットの法ベクトルをそれぞれ $\boldsymbol{h}_j$ と $\boldsymbol{h}_j^*$ $(j = 1, 2)$，また分離超平面を $\mathcal{H}_j$ と $\mathcal{H}_j^*$ のように区別する．さらに，階層性を明確にするため，出力ユニットの法ベクトル（分離超平面）を極性に応じて $\boldsymbol{s}$ と $\boldsymbol{s}^*$ ($\mathcal{S}$ と $\mathcal{S}^*$) のように表す．



(a) 分離超平面　(b) 両極型ネットの解領域　(c) 片極型ネットの解領域

図 3.2: XOR 問題の解となる中間層ユニットの分離超平面と法ベクトルによる解領域



(a) 分離超平面　(b) 両極型ネットの解領域　(c) 片極型ネットの解領域

図 3.3: XOR 問題の解となる出力層ユニットの分離超平面と法ベクトルによる解領域

3.3.1　両極型ネットの場合

入力パターンが $\boldsymbol{x}_1 = [0.5, 0.5]^T$, $\boldsymbol{x}_2 = [-0.5, 0.5]^T$, $\boldsymbol{x}_3 = [-0.5, -0.5]^T$, $\boldsymbol{x}_4 = [0.5, -0.5]^T$ で与えられる XOR 問題を $2-2-1$ の両極型ネットで解く場合を考える．この場合，$\boldsymbol{x}_k$ に対する目標出力 d_k $(k = 1, 2, 3, 4)$ は，$d_1 = d_3 = -1/2$, $d_2 = d_4 = 1/2$ となる．

中間層ユニットについては，$\boldsymbol{p}_n (= \boldsymbol{x}_n)$ の指す点を P_n $(n = 1, 2, 3, 4)$ として，これらを頂点とする正方形を入力パターン空間 X とすると，入力空間分割の観点から，XOR 問題の解となる $\mathcal{H}_1$ と $\mathcal{H}_2$ の対として次の 2 種類が存在することが知られている[49]．すなわち，図 3.2(a) に示すように，頂点 P_1 を頂点 $\mathrm{P}_2\mathrm{P}_4$ と分離する $\mathcal{H}_1$ と頂点 P_3 を頂点 $\mathrm{P}_2\mathrm{P}_4$ と分離する $\mathcal{H}_2$ の対 (Type I) と，頂点 P_2 を頂点 $\mathrm{P}_1\mathrm{P}_3$ と分離する $\mathcal{H}_1$ と頂点 P_4 を頂点 $\mathrm{P}_1\mathrm{P}_3$ と分離する $\mathcal{H}_2$ の対 (Type II) が存在する．

頂点 P_n を通る分離超平面に随伴する法ベクトルの軌跡は，原点から P_n までの線分を直径とする円

$$\Phi_n \ : \ \|\boldsymbol{h}_j - \boldsymbol{p}_n/2\|^2 = \|\boldsymbol{p}_n/2\|^2 \tag{3.3}$$

となる．Φ_n $(n = 1, 2, 3, 4)$ は，具体的に，それぞれ中心が $(1/4, 1/4)$, $(-1/4, 1/4)$, $(-1/4, -1/4)$, $(1/4, -1/4)$ で半径が $\sqrt{2}/4$ の円として与えられる．便宜のため，これらの円によって区分けされる 8 個の小領域を図 3.2(b) のように $A, B, \cdots, H$ と定める．

このとき，法ベクトル $\boldsymbol{h}_1(\boldsymbol{h}_2)$ が小領域 $A(E)$ にあれば，それに対応する分離超平面 $\mathcal{H}_1$ $(\mathcal{H}_2)$ は頂点 P_1 (P_3) と頂点 $\mathrm{P}_2\mathrm{P}_4$ を分離することが簡単な作図により確かめられる．したがって，$\boldsymbol{h}_1 \in A$

$= \Omega_1 - (\Omega_2 \cup \Omega_4)$ かつ $\boldsymbol{h}_2 \in E = \Omega_3 - (\Omega_2 \cup \Omega_4)$ ならば，$\boldsymbol{h}_1$ と $\boldsymbol{h}_2$ の対は Type I の解となって，その分布領域（解領域）は図 3.2(b) の斜線部となる．ここに，$\cup$ と $-$ は集合の和と差を表し，Ω_n $(= 1, 2, 3, 4)$ は

$$\Omega_n = \{\boldsymbol{h}_j \,|\, \|\boldsymbol{h}_j - \boldsymbol{p}_n/2\|^2 < \|\boldsymbol{p}_n/2\|^2\} \tag{3.4}$$

なる円板で定義される集合である．同様に，$\boldsymbol{h}_1 \in C = \Omega_2 - (\Omega_1 \cup \Omega_3)$ かつ $\boldsymbol{h}_2 \in G = \Omega_4 - (\Omega_1 \cup \Omega_3)$ ならば，$\boldsymbol{h}_1$ と $\boldsymbol{h}_2$ の対は Type II の解となり，解領域は図 3.2(b) の点表示された部分となる．

入力パターン空間は分離超平面を境に高出力域と低出力域に分かれ，どちらが高出力域になるかはバイアスの符号で一意に決まる．すなわち，$\mathcal{H}_j$ からみて原点側の領域は，$\boldsymbol{h}_j^T(\boldsymbol{x} - \boldsymbol{h}_j) = -b_j u_j / \|\boldsymbol{w}_j\|^2 < 0$ となって，バイアスが正（負）ならば高（低）出力域となる．これを念頭に図 3.2(a) を見てみると，例えば分離超平面が $\boldsymbol{h}_1 \in A$, $\boldsymbol{h}_2 \in E$ なる法ベクトルで与えられる Type I の場合，入力パターン $\boldsymbol{x}_k$ に対する出力 $\boldsymbol{y}_k = [y_{k1}, y_{k2}]^T$ が表 3.3 のように得られる．そして，Type II の $\boldsymbol{h}_1 \in C$, $\boldsymbol{h}_2 \in G$ なる場合の出力についても，表 3.2 と同様な結果が得られる．さらに，以上より，中間層からの出力空間，つまり出力ユニットに対する入力空間 Y は，X と同様，頂点を $Q_1Q_2Q_3Q_4$ とする正方形となることが分かる（図 3.3(a)）．

以下では，中間層ユニットが $\boldsymbol{h}_1 \in A$, $\boldsymbol{h}_2 \in E$ なる Type I の場合の出力ユニットの解について考える．表 3.2 の結果から，$b_1 > 0$, $b_2 > 0$ のとき，$\boldsymbol{y}_2$ と $\boldsymbol{y}_4$ は Y の第 1 象限の頂点 Q_1 に，$\boldsymbol{y}_1$ と $\boldsymbol{y}_3$ はそれぞれ第 2 および第 4 象限の頂点 Q_2 と Q_4 に配置される．したがって，出力ユニットの解は，頂点 Q_1 と頂点 Q_2Q_4 を分ける Type i の分離超平面 $\boldsymbol{S}$ で与えられることになる．さらに，他の b_1 と b_2 の組合せについても調べてみると，$\boldsymbol{h}_1 \in A$, $\boldsymbol{h}_2 \in E$ の場合，解となる $\boldsymbol{S}$ は図 3.3(a) に示すように Type i ～ iv の 4 種類が存在し，出力ユニットに対する解領域は図 3.3(b) に示す小領域 A', C', E', G' で与えられることが分かる．つまり，この場合，XOR 問題に対する解領域は，表 2 に示すように，中間層ユニットに対する小領域 A, E と，出力ユニットに対する A', C', E', G' のいづれか 1 つの小領域との組合せで与えられる．

また，中間層ユニットの法ベクトルが Type II の $\boldsymbol{h}_1 \in C$, $\boldsymbol{h}_2 \in G$ なる場合についても，同様の議論が展開できて，その解領域は表 3.2 のようにまとめられる．そして，表 3.2 の結果から，XOR 問題に対して計 8 種の解が存在することが分かる．

表 3.2: $\boldsymbol{h}_1 \in A$, $\boldsymbol{h}_2 \in E$ のときの中間層ユニット出力

b_1	b_2	$\boldsymbol{y}_1$	$\boldsymbol{y}_2$	$\boldsymbol{y}_3$	$\boldsymbol{y}_4$
+	+	-.5 .5	.5 .5	.5 -.5	.5 .5
+	−	-.5 -.5	.5 -.5	.5 .5	.5 -.5
−	+	.5 .5	-.5 .5	-.5 -.5	-.5 .5
−	−	.5 -.5	-.5 -.5	-.5 .5	-.5 -.5

表 3.3: XOR 問題に対する解領域

		両極型ネット			片極型ネット		
b_1	b_2	$\boldsymbol{h}_1$	$\boldsymbol{h}_2$	$\boldsymbol{s}$	$\boldsymbol{h}_1^*$	$\boldsymbol{h}_2^*$	$\boldsymbol{s}^*$
+	+	A	E	A'	A^*	E^*	A'^*
−	+	A	E	C'	A^*	E^*	C'^*
−	−	A	E	E'	A^*	E^*	E'^*
+	−	A	E	G'	A^*	E^*	G'^*
+	+	C	G	A'	C^*	G^*	A'^*
−	+	C	G	C'	C^*	G^*	C'^*
−	−	C	G	E'	C^*	G^*	E'^*
+	−	C	G	G'	C^*	G^*	G'^*

3.3.2　片極型ネットの場合

入力パターンが $\boldsymbol{x}_1^* = [1,1]^T$, $\boldsymbol{x}_2^* = [0,1]^T$, $\boldsymbol{x}_3^* = [0,0]^T$, $\boldsymbol{x}_4^* = [1,0]^T$ で与えられる XOR 問題を $2-2-1$ の片極型ネットで解く場合を考える．このときの目標出力は，$d_1^* = d_3^* = 0$, $d_2^* = d_4^* = 1$ である．そして，$\boldsymbol{p}_n^* (= \boldsymbol{x}_n^*)$ の指す点を P_n^* $(n = 1,2,3,4)$ と記して，これらを頂点とする正方形を X^* とする．

この場合，$\boldsymbol{p}_n^* = \boldsymbol{p}_n + \mathbf{1}/2$ なる関係にあるから，X^* は X を $\mathbf{1}/2$ 平行移動したものとなる．ここに，$\mathbf{1}$ は $[1,1]^T$ なるベクトルである．また，任意の $\boldsymbol{p}$ について，$\boldsymbol{p}$ と $\boldsymbol{p}^*$ $(= \boldsymbol{p} + \mathbf{1}/2)$ の指す点を通る $\mathcal{H}_j$ と $\mathcal{H}_j^*$ は，その法ベクトル $\boldsymbol{h}_j$ と $\boldsymbol{h}_j^*$ が

$$\boldsymbol{h}_j^* \;=\; \boldsymbol{h}_j + \left(\frac{\mathbf{1}^T \boldsymbol{h}_j}{2\|\boldsymbol{h}_j\|^2} \right) \boldsymbol{h}_j \tag{3.5}$$

なる関係にあれば，平行となる[48]．以上のことは，$\mathcal{H}_j$ に随伴する法ベクトル $\boldsymbol{h}_j$ を (3.5) 式で写

像して得られる $\boldsymbol{h}_j^*$ の指す分離超平面 $\mathcal{H}_j^*$ による X^* の分割は，$\mathcal{H}_j$ による X の分割と幾何学的に等価となることを意味している．したがって，前述の $A, B, \cdots, H$ の (3.5) 式による写像を $A^*, B^*, \cdots, H^*$ とするとき，片極型ネットの中間層ユニットに対する解領域は，A^*, E^* (Type I) および C^*, G^* (Type II) となるはずである．

以下では，$A^*, B^*, \cdots, H^*$ を求めるため，Φ_n の (3.5) 式による非線形写像がどのようになるかを調べる．先ず，$\boldsymbol{h}_j \neq \boldsymbol{0}$ の場合，(3.3) 式の Φ_n と等価な関係 $\|\boldsymbol{h}_j\|^2 = \boldsymbol{p}_n^T \boldsymbol{h}_j$ を (3.5) 式に代入すると，$\boldsymbol{h}_j^* = (\boldsymbol{p}_n^{*T}\boldsymbol{h}_j / \boldsymbol{p}_n^T \boldsymbol{h}_j)\boldsymbol{h}_j$ となって，$\|\boldsymbol{h}_j^*\|^2 = \boldsymbol{p}_n^{*T}\boldsymbol{h}_j^*$ なる関係が導かれる．これは，$\boldsymbol{p}_n^*$ の指す点 P_n^* を通る分離超平面 $\mathcal{H}_j^*$ に随伴する法ベクトル $\boldsymbol{h}_j^*$ の円軌跡

$$\Phi_n^* : \|\boldsymbol{h}_j^* - \boldsymbol{p}_n^*/2\|^2 = \|\boldsymbol{p}_n^*/2\|^2 \tag{3.6}$$

に他ならない．具体的には，Φ_1^* が中心を $(1/2, 1/2)$ とする半径 $\sqrt{2}/2$ の円，また Φ_2^* と Φ_4^* はそれぞれ中心が $(0, 1/2)$ と $(1/2, 0)$ の半径 $1/2$ の円となる．しかし，Φ_3^* は原点に縮退したゼロベクトルとなる．一方，$\boldsymbol{h}_j = \boldsymbol{0}$ の写像，つまり，X の中心（R^2 の原点）を通る $\mathcal{H}_j$ と幾何学的に等価な形で X^* の中心（$\boldsymbol{1}/2$）を通る $\mathcal{H}_j^*$ に随伴する $\boldsymbol{h}_j^*$ の軌跡は，$(1/4, 1/4)$ を中心とする半径 $\sqrt{2}/4$ の円となる．したがって，この円を Φ_0^* と表記するとき，Φ_n の (3.5) 式による非線形写像は，Φ_n^* と Φ_0^* の 2 つの円の周上の点として与えられることになる．

以上の写像に関する結果から，$A^*, B^*, \cdots, H^*$ の境界が特定できて，その内部領域が決まる．その結果，片極型ネットの XOR 問題に対する中間層ユニットの解領域は，図 3.2(c) の斜線部 (A^*, E^*) と点表示部 (C^*, G^*) になることが分かる．ここに，$A^* = \Omega_1^* - (\Omega_2^* \cup \Omega_4^*)$，$E^* = \Omega_2^* \cap \Omega_4^*$，$C^* = \Omega_2^* - (\Omega_1^* \cup \Omega_4^*)$，$G^* = \Omega_4^* - (\Omega_1^* \cup \Omega_4^*)$ で，Ω_n^* は

$$\Omega_n^* = \{\boldsymbol{h}_j^* \mid \|\boldsymbol{h}_j^* - \boldsymbol{p}_n^*/2\|^2 < \|\boldsymbol{p}_n^*/2\|^2\} \tag{3.7}$$

なる法ベクトルの集合である．

出力ユニットの分離超平面 $\mathcal{S}^*$ や法ベクトル $\boldsymbol{s}^*$ についても，同様な議論を展開することにより，片極型ネットによる XOR 問題の解領域が表 3.2 のように得られる．例えば $\boldsymbol{h}_1^* \in A^*$, $\boldsymbol{h}_2^* \in E^*$, $b_1 > 0$, $b_2 > 0$ の場合，XOR 問題に対する片極型ネットの解領域は，中間層ユニットの法ベクトル $\boldsymbol{h}_j^*$ の属する小領域 A^*, E^*（図 3.2(c)）と出力ユニットの法ベクトル $\boldsymbol{s}^*$ の属する小領域 A'^*（図 3.3(c)）の組合せで与えられる．

3.3.3　解領域の極性による相違

以上をもとに，図3.2(b)(c)および図3.3(b)(c)において対応する小領域を比較すると，XOR問題に対する解領域は両極型と片極型ネットとで大きく異なることが分かる．そして，この相違を(2.17)式の法ベクトルの定義と照合することにより，荷重およびバイアス解として，例えば，これらがすべて同一符号となる解は，両極型では存在するが，片極型では存在しないことが明らかとなる．さらに，これらの結果から，一般的な課題を通常のようにユニットがシグモイド特性の多層ネットでBP学習させる場合，その解領域やアトラクタはネットの極性に依存して異なると推察される．

尚，分離超平面が入力空間内を通ることを解の必要条件とすれば，両極型のとき $\Omega = \bigcup_{n=1}^{4} \Omega_n$，片極型のとき $\Omega^* = \bigcup_{n=1}^{4} \Omega_n^*$ が必要条件となる．また，両極型ネットに入力パターンとして $\boldsymbol{x}_k$ の代わりに $\boldsymbol{x}_k^*$ を提示したときの中間層ユニットの解領域は，図3.2(c)に示す小領域 A^*, E^*, C^*, G^* で与えられる．

3.4　極性による収束能力の違い

ここでは，極性による収束能力の違いをパリティ問題とランダムマッピング問題を対象としたシミュレーション実験により検証する．ネットワークは3層構成 $(m-n-k)$ のものとして，荷重の初期値は，幅が $(-\gamma, \gamma)$ の一様乱数で $\gamma = 0.01, 0.1, 0.5, 1.0, 1.5, \cdots, 6.0$ として与えた．また学習は，慣性係数を $\alpha = 0.9$ として，学習係数が $\eta = 0.2, 0.6, 1$ の場合を試みた．そして，収束判定基準は，全パターンに対して，個々の出力ユニットにおける誤差のすべてが，$|\varepsilon_j| < \varepsilon^*$ $(j = 1, 2, \cdots, k)$ と ε^* 以内に収まることとし，パリティ問題においては $\varepsilon^* = 0.1$，またランダムマッピング問題においては $\varepsilon^* = 0.05$ とした．また，収束の評価は，50回の試行に基づいて行った．

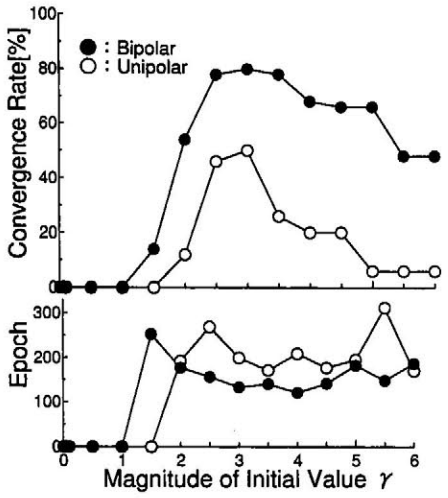


図 3.4: 6 次のパリティ問題の学習結果

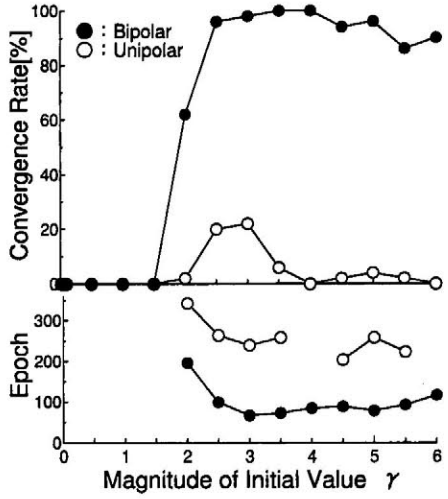


図 3.5: 8 次のパリティ問題の学習結果

3.4.1 パリティ問題の学習結果

まず，パリティ問題について，訓練パターンや慣性係数を前述のように与えて，アルゴリズムの最大繰返し回数を $m \leq 6$ のとき 30,000 回，$m \geq 7$ のとき 100,000 回として，シミュレーションを行った．ネットワークサイズを大きくしたときの学習例として，6 次と 8 次のパリティ問題をそれぞれ $6-9-1$ と $8-30-1$ のネットワークで学習させた結果 ($\eta = 0.6$) を図 3.4 と図 3.5 に示す．図において，横軸の "Magnitude of Initial Value γ" は初期値の大きさ γ，縦軸の "Convergence Rate" は収束率，"Epoch" は平均学習回数を表す．そして，片極型による結果を○，両極型による結果を▲で示し，収束率を実線，平均学習回数を破線で結んだ．また，活性領域幅 g_j^0 の平均 g については，$\gamma = 1$ のときの g を $g_{\gamma=1}$ と表記するとき，$m = 6, 8, 9, 30$ に対して，それぞれ $g_{\gamma=1} = 3.31, 2.81, 2.64, 1.40$ となり，その他の γ については $g_{\gamma=1}/\gamma$ であった．さらに，$\|\boldsymbol{h}_j^0\|$ の平均 h については，すべての γ に対して，$m = 6, 8, 9, 30$ のとき，それぞれ $h = 0.377, 0.318, 0.299, 0.160$ であった．

図 3.4 と図 3.5 の結果より，ネットワークサイズが大きい場合，良好な収束を与える初期値 γ の範囲は，サイズが小さいときの結果（表 3.1）と違って，片極型と両極型でほぼ共通していることが分かる．そして，表 3.1 に示すように，サイズが小さい場合の最高の収束率は，片極型と両極型でそれぞれ 0.1 と 1.0 と比較的小さな γ で得られたのに対して，ここでのようにネットワーク

サイズが大きい場合，最高の収束率を与える γ はいずれも 3 付近と大きくなっていることが読み取れる．さらに，収束率および平均学習回数ともに，両極型による結果が片極型に比べて優れており，良好な収束率を与える γ の範囲も両極型による方が広いことが分かる．他にも学習係数を $\eta = 0.2, 1.0$ として比較した結果，片極型による収束率が η の違いによる影響を受け易いのに対して，両極型による収束率は比較的 η の影響を受けにくいことや，極性による差は η が 0.2 と小さいほど少なく，η が 1.0 と大きいほど開くことが認められた[49]．

片極型の場合，何れのパリティ問題においても，収束が良好となる適切な初期荷重の範囲（等価的に，適切な活性領域幅）が存在することが，図 3.4 や図 3.5 の結果から読みとれる．しかし，両極型の場合，適切な荷重の下限は読みとれるが，その上限は明らかでない．そこで，両極型に対して，試行回数を１０回にして，γ が 7 以上の範囲について実験した．その結果，6 次のパリティ問題では，$\gamma = 8.0, 10.0, 12.0$ のとき，収束率はそれぞれ 40 %, 10 %, 0 % となり，また 8 次のパリティ問題では，$\gamma = 8.0, 10.0, 12.0$ のとき，収束率はそれぞれ 90 %, 30 %, 10 % となって，いずれも上限の存在を確認できた．また，他の学習係数についても同様に上限が確認された．

3.4.2　ランダムマッピング問題の学習結果

次に，ランダムマッピング問題については，片極型ネットワークの場合，幅が $(0,1)$ の一様乱数により生成される 5 個の乱数 r_i $(i = 1, 2, \cdots, 5)$ の組を 1 つのパターンとして，これを入力と教師として与え，全部で 9 パターンを学習させた．また，両極型ネットワークの場合，$r_i - 0.5$ なる乱数を用いた．そして，最大繰返し回数は 30,000 回とした．

図 3.6 に $5-5-5$ のネットワークで $\eta = 0.6$ として学習させたときの結果を示す．この図から，両極型による平均学習回数は片極型に比べて優れており，良好な収束率を与える γ の範囲も両極型による方が広いことが分かる．このことは，η を 0.2 や 1.0 としたときも同様であった．また，両極型による場合，学習係数の収束率に及ぼす影響は片極型に比べて少ないことが認められた．しかし，両者による収束率の差は，η が 0.2 と小さくなるほど開き，パリティ問題のときとは逆の傾向を示していた[49]．

さらに，他にも１０種類ほど異なる乱数パターンについて試行してみたが，いずれも図 3.6 と

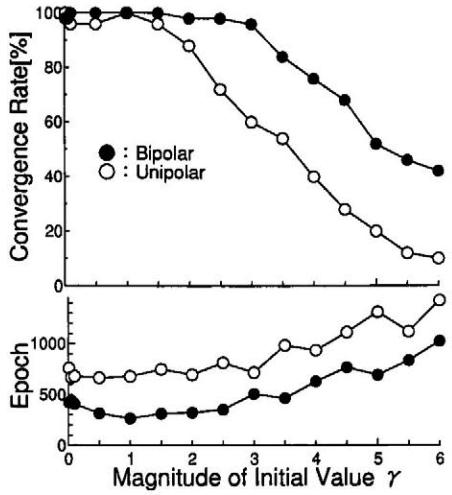


図 3.6: ランダムマッピング問題の学習結果

同様な結果が得られて，収束率・平均学習回数ともに両極型による方が優れていることが分かった．また，$5-7-5$ のネットワークで学習させた場合も図 3.6 と同様な結果であった．

図 3.6 に示した結果では，$\gamma = 0.01$ のときでも，両極型と片極型の双方でほぼ 100 % の収束率となっている．このときの $\|\boldsymbol{h}_j^0\|$ の平均は $h = 0.423$ と違和感のない値であるが，活性領域幅の平均と入力定義域サイズの比は $g/D = 165.6$ とかなり大きな値である．確かに，前述のように，活性領域幅には学習問題に応じて適切な範囲があるが，この値はあまりにも大きすぎる．そこで，乱数パターンに原因がないかどうかを調べるため，中間ユニットを 1 個少なくした $5-4-5$ のネットワークで学習させた結果，両極型と片極型の双方とも全く収束せず，乱数パターンには問題がないことが分かった．さらに，初期値を小さくしてみたら，さすがに $\gamma = 0$ では収束しなかったものの，$\gamma = 10^{-6}$ 程度までは，両極型・片極型ともに，$\gamma = 0.01$ のときとほぼ同じ収束率と平均学習回数を示すことが判明した．また，収束判定基準 ε^* を厳しくして，$10^{-6} \leq \gamma \leq 6$ の範囲を対象に調べた結果，$\varepsilon^* = 0.02$ のときでも，γ が小さい方での収束率の落ち込みは確認できなかったが，$\varepsilon^* = 0.01$ のとき，片極型では $\gamma = 10^{-3}$，また両極型では $\gamma = 0.1$ での収束率２０数パーセントをピークに，γ がそれより増大または減少するにつれて，収束率は低下することを確認した．そして，$\varepsilon^* = 0.03, 0.02, 0.01$ のときの収束率や平均学習回数は，$\gamma \leq 0.001$ で片極型の方が，また $\gamma \geq 0.1$ で両極型の方が優位となることを確認した．

以上により，ネットワークサイズが大きい場合，片極型に比べて，両極型による学習は，ゼロの近傍を除いて，より広い範囲の初期値に対して良好な収束を与えることを確認した．また，両極型による学習は，学習係数の影響を比較的受けにくいことが分かった．

3.5　収束挙動に関する解析

2.5節で述べたように両極型ネットと片極型ネットの分割は異なる．そこで，分離超平面による各層での入力定義域分割が幾何学的に等しく，入力定義域に占める活性領域の幅も同じくすれば，提示パターンに対する誤差特性は，片極型ネットワークか両極型ネットワークかによらず同一になることを述べる．このことは，極性によらず解領域の大きさが同じになることを意味している．しかし，分割や活性領域を同じにして，学習を開始させたとしても，その後の収束挙動が一致するとは限らないことを述べる．

3.5.1　入力定義域の等価な分割

ここでは，両極型ネットと片極型ネットのそれぞれの入力定義域の分割が等価となるための条件について考える．以降では，あるユニットの結合荷重ベクトルとバイアス荷重の表記を，片極型のとき $\boldsymbol{w}_U$ と b_U，また両極型のとき $\boldsymbol{w}_B$ と b_B と区別する．この場合，活性幅の定義 (2.18) 式より，片極型と両極型の荷重を

$$\boldsymbol{w}_U \;=\; \boldsymbol{w}_B \tag{3.8}$$

と等しくすれば，活性領域の幅は極性によらず同じになることがわかる．

この状況下で，入力定義域の分割が片極型ネットワークと両極型ネットワークとで幾何学的に等価となる例を図3.7に示す．図において，H_B と H_U（X_B と X_U）はそれぞれ両極型ネットワークのときと片極型ネットワークのときの分離超平面（入力定義域），P（Q）は H_B（H_U）と原点 O からそれへの垂線との交点，R は X_U の中心，S は 線分 OQ と R からそれへの垂線との交点である．この場合，入力定義域は幾何学的に等価に分割されていることから，ベクトル $\overrightarrow{\mathrm{OP}}$ は $\overrightarrow{\mathrm{SQ}}$ と等しく，$\overrightarrow{\mathrm{OQ}}$ は $\overrightarrow{\mathrm{OS}}$ と $\overrightarrow{\mathrm{SQ}}$ の和で表されることがわかる．したがって，前節の定義 (2.17)

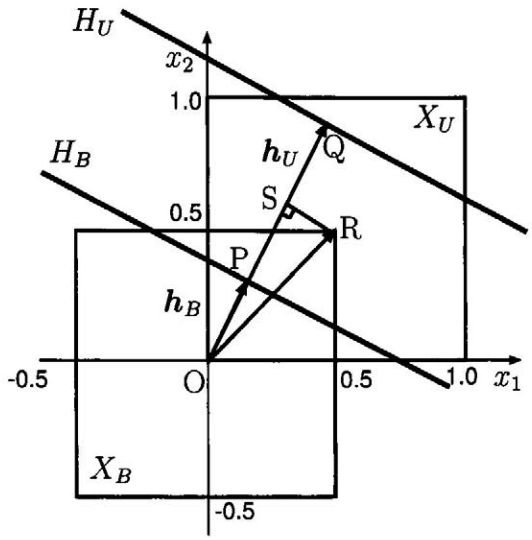


図 3.7: 入力定義域の等価な分割

式に準じて，ベクトル $\overrightarrow{\mathrm{OP}}$ と $\overrightarrow{\mathrm{OQ}}$ をそれぞれ法ベクトル $\boldsymbol{h}_B$ と $\boldsymbol{h}_U$ と記述するとき，$\boldsymbol{h}_U = \overrightarrow{\mathrm{OS}} + \boldsymbol{h}_B$ となる．また，$\overrightarrow{\mathrm{OS}} = \mathbf{1}^T \boldsymbol{h}_B \boldsymbol{h}_B / (2\|\boldsymbol{h}_B\|^2)$ となることより，

$$\boldsymbol{h}_U = (1 + \frac{\mathbf{1}^T \boldsymbol{h}_B}{2\|\boldsymbol{h}_B\|^2})\boldsymbol{h}_B \tag{3.9}$$

なる関係を得る．ここに，$\mathbf{1}$ は $[1, 1, \cdots, 1]^T$ のように要素がすべて 1 となるベクトルである．さらに，(3.9) 式 に (3.8) 式 を代入すると，片極型のときのバイアス荷重 b_U と両極型のバイアス荷重 b_B の間に，

$$b_U = b_B - \frac{1}{2}\sum_{i=1}^{m} w_i^B \tag{3.10}$$

なる関係を得る．ここに，w_i^B は $\boldsymbol{w}_B$ の第 i 番目の要素である．

したがって，両極型ネットワークに入力パターン $\boldsymbol{x}$ を提示し，片極型ネットワークには $\boldsymbol{x} + 1/2$ なる入力パターンを提示するとき，第 1 中間ユニットの結合荷重とバイアス荷重をそれぞれ (3.8) 式や (3.10) 式 のように与えれば，両極型と片極型のネットワークに対する入力パターン空間 $P_B = \{\boldsymbol{x}\}$ と $P_U = \{\boldsymbol{x} + \mathbf{1}/2\}$ の分割は幾何学的に等価で，活性領域も幾何学的に等価な領域を占めることになる[49]．

3.5.2　誤差特性と解領域の大きさ

いま，両極型ネットワークに入力パターン $\boldsymbol{x}$ を提示するとき，片極型ネットワークには $\boldsymbol{x}+\mathbf{1}/2$ なる入力パターンを提示することにする．この場合，第1中間ユニットの荷重を (3.8)(3.10) 式のように与えれば，両極型と片極型のネットワークに対する入力パターン空間 $P_B = \{\boldsymbol{x}\}$ と $P_U = \{\boldsymbol{x}+\mathbf{1}/2\}$ の分割は幾何学的に等価で，活性領域も幾何学的に等価な領域を占めることになる．

以下では，すべての層で，荷重を (3.8)(3.10) 式のように与えた場合，両極型ネットワークに対して $(\boldsymbol{x}, \boldsymbol{d})$ なる入力パターン・教師対を提示したときの誤差と，片極型ネットワークに対して $(\boldsymbol{x}+\mathbf{1}/2, \boldsymbol{d}+\mathbf{1}/2)$ なる入力パターン・教師対を提示したときの誤差は等しくなることを述べる．すなわち，この場合，すべての層で，入力定義域の幾何学的分割や活性領域の幅は両極型と片極型ネットワークで等しいことから，両極型のとき，$\boldsymbol{x}$ なる入力パターンに対する第1中間層の出力が $\boldsymbol{y}$ となることは，片極型のとき，$\boldsymbol{x}+\mathbf{1}/2$ なる入力パターンに対する第1中間層の出力が $\boldsymbol{y}+\mathbf{1}/2$ となることを意味する．したがって，第2中間層に対する入力は，両極型のとき $\boldsymbol{y}$，片極型のとき $\boldsymbol{y}+\mathbf{1}/2$ となり，第3中間ユニットによる入力定義域の幾何学的分割や活性領域の幅も両極型と片極型で等しくなる．そして，同じことが最終の出力層まで繰り返されるから，両極型ネットワークの出力を $\boldsymbol{z}$ とするとき，片極型ネットワークの出力は $\boldsymbol{z}+\mathbf{1}/2$ となる．このとき，教師を，出力ユニットが両極型のとき $\boldsymbol{d}$，片極型のとき $\boldsymbol{d}+\mathbf{1}/2$ と与えれば，提示入力パターンに対する誤差は極性によらず同じく $\boldsymbol{\epsilon} = \boldsymbol{d} - \boldsymbol{z}$ となる．それゆえ，両極型ネットワークによる誤差と，片極型のネットワークによる誤差は等しくなる．

以上のことは，$\{\boldsymbol{w}_B^*, b_B^*\}$ が両極型ネットワークに対する収束解ならば，$\{\boldsymbol{w}_U, b_U \mid \boldsymbol{w}_U = \boldsymbol{w}_B^*, b_U = b_B^* - \mathbf{1}^T\boldsymbol{w}_B^*/2\}$ は片極型のネットワークに対する収束解となることを示唆している．そして，これらの収束解は1対1に対応する．したがって，両極型と片極型ネットワークは同じ大きさの解領域をもつことになる．

また，片極型と両極型ネットワークの両方に共通の入力パターン $\boldsymbol{x}$ を提示する場合も，解領域は同じ大きさとなる．すなわち，この場合，第1中間ユニットの荷重が両極型と片極型ネットワークで同じならば，入力パターン空間における分割や活性領域は双方で同じとなる．そして，第2中間層以降のユニットの荷重を (3.8)(3.10) 式のように与えた場合，入力パターンに対する誤差特

性は極性によらず同一となる．それゆえ，解領域の大きさは同じとなる．

以下では，議論の便宜を計るため，1つの層は同じ極性のユニットから成るものとして，ネットワーク構成とそれへのパターンの提示を簡便に表記する．すなわち，両極型と片極型のユニットについて，荷重がそれぞれ $\boldsymbol{w}_U = \boldsymbol{w}_B$ や $b_U = b_B$ と等しく与えられるとき，両極型のものをB，片極型のものをUと表し，片極型のうち，バイアス荷重が(3.10)式の関係で与えられるものをU‡と表す．このとき，上述のネットワークは，入力定義域の幾何学的分割や活性領域の幅が等価で，同じ誤差特性や解領域の大きさをもつグループとして，$\boldsymbol{x}$ - B - ⋯ - B - $\boldsymbol{d}$，$\boldsymbol{x}$† - U‡ - ⋯ - U‡ - $\boldsymbol{d}$†，$\boldsymbol{x}$ - U - U‡ - ⋯ - U‡ - $\boldsymbol{d}$† のようにまとめられる．ここに，$\boldsymbol{x}$ - B - ⋯ - B - $\boldsymbol{d}$ は両極型ネットワークに入力パターン と教師の対 $\{(\boldsymbol{x},\boldsymbol{d})\}$ が提示されることを表し，$\boldsymbol{x}$† は，入力パターンが $\boldsymbol{x}\dagger = \boldsymbol{x} + \mathbf{1}/2$ のように，また，$\boldsymbol{d}$† は，教師が $\boldsymbol{d}\dagger = \boldsymbol{d} + \mathbf{1}/2$ のように提示されることを表す．

3.5.3 収束挙動

前節で述べたように，すべての層において，入力定義域の幾何学的な分割や活性領域の幅が同じ場合，ネットワークの誤差特性は極性によらず同一となる．しかし，このような状況下で学習を開始したとしても，その後の収束挙動は必ずしも一致するとは限らない．以下では，このことについて考察する．

まず，上述の3つのネットワーク $\boldsymbol{x}$ - B - ⋯ - B - $\boldsymbol{d}$，$\boldsymbol{x}$† - U‡ - ⋯ - U‡ - $\boldsymbol{d}$†，$\boldsymbol{x}$ - U - U‡ - ⋯ - U‡ - $\boldsymbol{d}$† による学習について考える．これらのネットワークは，前節で述べたように，最初の入力パターン（第1中間層が両極型のとき $\boldsymbol{x}$，片極型のとき $\boldsymbol{x} + \mathbf{1}/2$）と教師（出力層が両極型のとき $\boldsymbol{d}$，片極型のとき $\boldsymbol{d} + \mathbf{1}/2$）の提示に対して，同じ誤差特性を示す．さらに，これらの出力ユニットの活性領域はその入力定義域において幾何学的に等価な位置を占めることから，シグモイド関数の微分 $f'(\cdot)$ も同じになる．すなわち，出力ユニットについて，両極型の場合の出力を z_B，片極型の場合の出力を z_U とすれば，両者は $z_U = z_B + 1/2$ なる関係にある．したがって，シグモイド関数の微分については，その表現形式こそ

$$f'_B(\cdot) \;=\; (0.5 + z_B)(0.5 - z_B) \tag{3.11}$$

$$f'_U(\cdot) \;=\; z_U(1-z_U) \tag{3.12}$$

と極性によって異なるが，値そのものは同じとなる．それゆえ，逆伝搬誤差

$$\delta_k = f'(\cdot)\varepsilon_k \tag{3.13}$$

も同じ値となる．したがって，ＢＰ学習による出力ユニットの結合荷重 w_{kj} の更新は，そのユニットへの入力を y_j とするとき，片極型か両極型によらず同一の形式で，

$$\Delta w_{kj} \;=\; \eta\delta_k y_j + \alpha\Delta w_{kj}(old) \tag{3.14}$$

のように記述される．ここに，η は学習係数，α は慣性係数である．

しかし，出力ユニットへの入力 y_j については，下位層が両極型のときの値を y_B とすると，下位層が片極型の場合，$y_B + 1/2$ なる値をとることになるので，(3.14) 式の右辺の第１項目の値は異なる．したがって，Δw_{kj} や更新後の結合荷重 $w_{kj} + \Delta w_{kj}$ の値は，下位層ユニットの極性によって異なる．同様に下位層での荷重更新も異なる．それゆえ，上記の３つのネットワークによる学習は，以降での提示パターンに対する誤差がもはや同一とならないため，異なる収束挙動を呈すことになる．

また，すべての中間層を両極型のユニットにして出力層のみを片極型のユニットにしたネットワーク $\boldsymbol{x}$ - B - ･･･ - B - U - $\boldsymbol{d}$† と両極型ネットワーク $\boldsymbol{x}$ - B - ･･･ - B - B - $\boldsymbol{d}$ による場合，収束挙動は一致することが以下の議論からわかる．

つぎに，すべての中間層を両極型のユニットにして出力層のみを片極型のユニットにしたネットワーク $\boldsymbol{x}$ - B - ･･･ - B - U - $\boldsymbol{d}$† と両極型ネットワーク $\boldsymbol{x}$ - B - ･･･ - B - B - $\boldsymbol{d}$ による学習について考える．この場合，両者の荷重の初期値は同じであるから，以下のように両者による収束挙動は一致することが導かれる．すなわち，この場合，入力パターン $\boldsymbol{x}$ の提示に対する出力ユニットへの入力 y_j は，双方のネットワークで同じ値となって，出力層に対する入力定義域は一致する．また，出力ユニットでの荷重も同じであるから，出力ユニットのなす分離超平面や活性領域はその入力定義域で全く同じ場所を占めることになる．したがって，入力パターン $\boldsymbol{x}$ と教師（出力ユニットが両極型のとき $\boldsymbol{d}$，片極型のとき $\boldsymbol{d}+\mathbf{1}/2$ ）の提示に対する誤差は双方のネットワークで等しい．それゆえ，(3.14) 式による更新値は同じものとなる．さらに，出力層のバイアス荷重更

新や，下位層ユニットの荷重の更新についても，双方で同じになる．したがって，2つのネットワークによる学習は，パターンが同じ順番で提示されるとき，同一の収束挙動を示すことになる．

以上のことから，両極型ネットワークとの対比で，誤差特性のみが等しくて収束挙動の異なるグループを

$\boldsymbol{x}$ - B - ⋯ - B - $\boldsymbol{d}$

$\boldsymbol{x}$† - U‡ - ⋯ - U‡ - $\boldsymbol{d}$†

$\boldsymbol{x}$ - U - U‡ - ⋯ - U‡ - $\boldsymbol{d}$†

のように，また，誤差特性のみならず収束挙動も等しいグループを

$\boldsymbol{x}$ - B - ⋯ - B - B - $\boldsymbol{d}$

$\boldsymbol{x}$ - B - ⋯ - B - U - $\boldsymbol{d}$†

のようにまとめることができる．

また，片極型ネットワークとの対比では，誤差特性のみが等しくて収束挙動の異なるネットワークを

$\boldsymbol{x}$† - U - ⋯ - U - $\boldsymbol{d}$†

$\boldsymbol{x}$ - B‡ - ⋯ - B‡ - $\boldsymbol{d}$

$\boldsymbol{x}$ - B - B‡ - ⋯ - B‡ - $\boldsymbol{d}$

のように，収束挙動も等しいネットワークを

$\boldsymbol{x}$† - U - ⋯ - U - U - $\boldsymbol{d}$†

$\boldsymbol{x}$† - U - ⋯ - U - B - $\boldsymbol{d}$

のようにまとめられる．ここに，B‡ は，そのバイアス荷重が

$$b_B = b_U + \frac{1}{2}\sum_{i=1}^{m} w_i^U$$

と与えられるときの両極型ユニットを指す．

3.6　シミュレーション

ここでは，ネットワークが3層で構成される場合を考え，両極型ネットワーク $\boldsymbol{x}$ - B - B - $\boldsymbol{d}$ と，入力定義域の幾何学的な分割や活性領域の幅が等しいネットワーク $\boldsymbol{x}$† - U‡ - U‡ - $\boldsymbol{d}$†，$\boldsymbol{x}$ - U - U‡ - $\boldsymbol{d}$†，$\boldsymbol{x}$ - B - U - $\boldsymbol{d}$† について，パリティ問題やランダムマッピング問題を対象に誤差特性や収束挙動をシミュレーションにより調べる．また，比較のため，片極型ネットワーク $\boldsymbol{x}$† - U - U - $\boldsymbol{d}$† についても調べる．

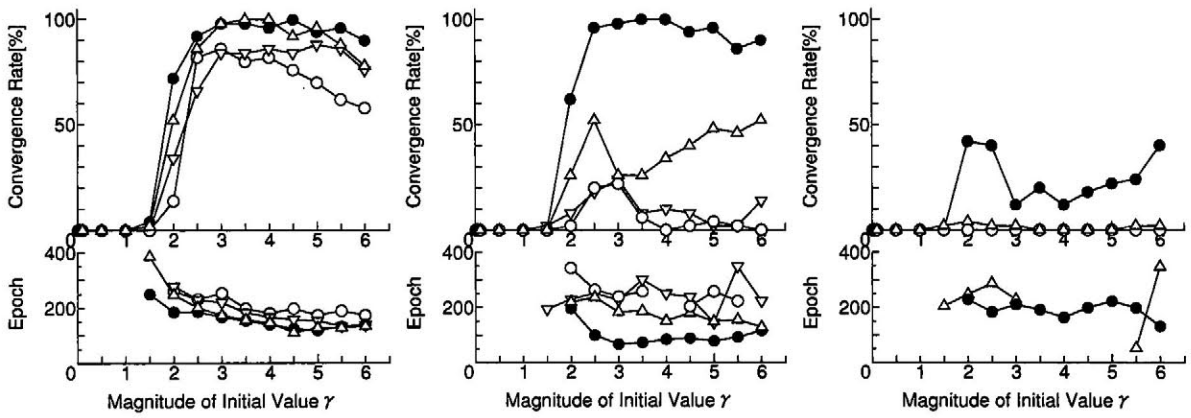
シミュレーションにおいて，荷重の初期値は，片極型ネットワークと両極型ネットワークの場合，幅が $(-\gamma, \gamma)$ の一様乱数で与え，その他のネットワークの場合，入力定義域の幾何学的分割や活性領域の幅が両極型ネットワークのときと等しくなるように与えた．また学習は，慣性係数を $\alpha = 0.9$ として，学習係数が $\eta = 0.2, 0.6, 1$ のとき，$\gamma = 0.01, 0.1, 0.5, 1.0, 1.5, \cdots, 6.0$ のそれぞれについて，50 回の乱数試行を試みた．そして，収束判定基準は，全パターンに対して，個々の出力ユニット $(k = 1, 2, \cdots, K)$ における誤差のすべてが $|\varepsilon_k| < \varepsilon^*$ に収まることとし，パリティ問題においては $\varepsilon^* = 0.1$，またランダムマッピング問題においては $\varepsilon^* = 0.05$ とした．

3.6.1　パリティ問題

入力パターンが $\boldsymbol{x}$ と表記されるとき，その要素を ±0.5，入力パターンが $\boldsymbol{x}$† と表記されるとき，その要素を 0 か 1 で与えた．この場合，分割が幾何学的に等価な3つのネットワークについては，中間ユニットによる分離超平面が，ユニットへの入力本数が増えるにつれて入力パターン空間の中心に漸近することになるが，片極型ネットワークについては，入力パターン空間の頂点に漸近することに注意されたい．さらに，教師の提示については，$\boldsymbol{d}$ のとき ±0.4，$\boldsymbol{d}$† のとき 0.1 か 0.9 で与えた．そして，パターンの提示順序は，すべてのネットワークで同一とした．

以上の実験条件のもとで，まず，中間層が2個のネットワーク（2 - 2 - 1）で，XOR問題の4個の提示パターンに対する誤差特性を調べた．その結果，片極型ネットワークを除いて，入力定義域の幾何学的な分割や活性領域の幅が等しい4つのネットワークは，いずれも同じ誤差特性を示すことが確かめられた．

つぎに，XOR 問題を 2 – 2 – 1 のネットワークで，繰返し回数の最大を 30,000 回として，学習させた．このときの収束に至るまでの誤差ならびに荷重の推定値を調べたところ，$\boldsymbol{x}$ - B - U - $\boldsymbol{d}\dagger$ と両極型ネットワークによる結果は全く同じ推移経過を示し，両者による収束挙動は一致することが確かめられた．また，その他のネットワークについては，収束挙動が両極型ネットワークと一致しないことが確かめられた．ただし，平均的な収束挙動を表す収束率（すべての提示パターンに対する誤差の絶対値が $\gamma = 0.1$ 未満に収まったときの試行を収束試行として，収束試行数を総試行数５０回で割った値）や平均学習回数（収束試行における平均の繰返し回数を提示パターン数で割った値）を調べたところ，良好な収束率の得られる初期値の大きさ γ が，両極型ネットワークや $\boldsymbol{x}$ - B - U - $\boldsymbol{d}\dagger$ では相対的に大きくなる傾向にあるのに対して，その他のネットワークでは小さくなる傾向にあるという点を除いて，これらによる平均的な収束挙動は大差ないことがわかった．



(a) $\eta = 0.2$　(b) $\eta = 0.6$　(c) $\eta = 1.0$

図 3.8: 8 次のパリティ問題による収束挙動の違い

そこで，サイズが大きなネットワークの場合，収束挙動がどの程度異なるかを調べるため，8 次のパリティ問題を 8 – 30 – 1 のネットワークで，最大の繰り返し回数を 100,000 回として，学習させた．このときの結果を図 3.8 に示す．図において，(a), (b), (c) はそれぞれ学習係数を $\eta =$ 0.2, 0.6, 1 としたときの結果で，横軸の "Magnitude of Inital Value" は初期値の大きさ γ，縦軸の "Convergernce Rate" は収束率，"Epoch" は平均学習回数を表す．また，●印は両極型ネット

ワーク $\boldsymbol{x}$ - B - B - $\boldsymbol{d}$ による結果，△と▽印はそれぞれ $\boldsymbol{x}$† - U‡ - U‡ - $\boldsymbol{d}$† と $\boldsymbol{x}$ - U - U‡ - $\boldsymbol{d}$† による結果，さらに○印は片極型ネットワーク $\boldsymbol{x}$† - U - U - $\boldsymbol{d}$† による結果である．

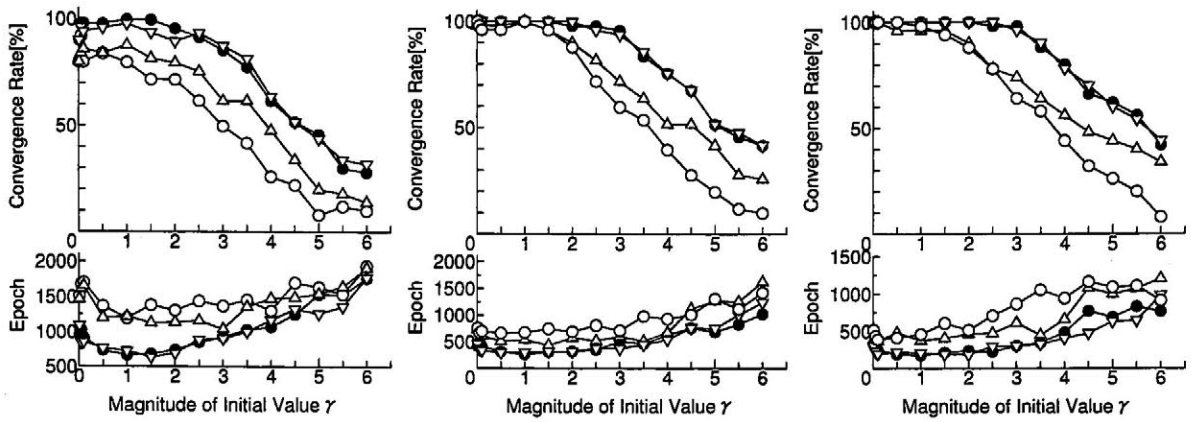
まず，$\boldsymbol{x}$† - U‡ - U‡ - $\boldsymbol{d}$† による結果をみてみると，その収束率は，両極型ネットワークによるものに比べて，学習係数が $\eta = 0.2$ と小さい場合，ほぼ互角であるが，学習係数が $\eta = 0.6, 1.0$ と大きくなるにつれて悪くなっている．同様なことが，平均学習回数についても云える．また，$\boldsymbol{x}$ - U - U‡ - $\boldsymbol{d}$† による場合，収束率・平均学習ともに，両極型ネットワークや $\boldsymbol{x}$† - U‡ - U‡ - $\boldsymbol{d}$† に比べても劣り，その結果は，むしろ，片極型ネットワーク $\boldsymbol{x}$† - U - U - $\boldsymbol{d}$† によるものに近い．

また，$\boldsymbol{x}$ - B - U - $\boldsymbol{d}$† と両極型ネットワークの収束挙動が一致することは，前章で証明し，前述のＸＯＲ問題でも検証した通りである．しかし，収束挙動が原理的に一致するはずのネットワークでも，上述の８次のパリティ問題に対するシミュレーションにおいて，収束挙動が厳密に一致しない場合が幾つかみられた．この相違は，以下のように計算機の数値計算精度に起因している．すなわち，２つのネットワークについて，荷重の推定値の推移経過を調べたところ，学習開始からしばらくの間，両者全く同じ値を示すものの，繰り返し計算の数百から数千回目あたりから，両者の間に 10^{-4} 程度のずれが生じて，それが徐々に大きくなって収束解やそれに至るまでの学習回数は違ったものになることが判明した．そして，ずれの発生は，出力ユニットの極性の違いにより，ユニット出力 ((2.3)(2.4) 式) や逆伝搬誤差におけるシグモイド関数の微分 ((3.11)(3.12) 式) 等の計算に精度の影響が現れたためと考えられる．そのため，８次のパリティ問題の場合，平均的な挙動を表す収束率や平均学習回数には，両者の間で僅かながらずれがあった．

3.6.2　ランダムマッピング問題

ランダムマッピング問題について，$5-5-5$ のネットワークで最大繰返し回数を 30,000 回として，学習させた．そして，入力パターンや教師については，幅が $(-0.5, 0.5)$ の一様乱数により生成される５個の乱数 r_i $(i = 1, 2, \cdots, 5)$ の組を１つのパターンとして全部で９パターンを用意し，$\boldsymbol{x} = \boldsymbol{d} = [r_1, r_2, \cdots, r_5]$ や $\boldsymbol{x}\dagger = \boldsymbol{d}\dagger = [r_1 + 0.5, r_2 + 0.5, \cdots, r_5 + 0.5]$ のように与えた．

学習係数が $\eta = 0.2, 0.6, 1.0$ のときの結果をそれぞれ図 3.9 の (a), (b), (c) に示す．図の見方は，前述のパリティ問題のときと同様である．両極型ネットワーク（●）と $\boldsymbol{x}$ - U - U‡ - $\boldsymbol{d}$† （▽）に



(a) $\eta = 0.2$ (b) $\eta = 0.6$ (c) $\eta = 1.0$

図 3.9: ランダムマッピング問題による収束挙動の違い

よる収束率と平均学習回数は，いずれの η についても，同様な結果となっていることが分かる．しかし，これに比べて，$\boldsymbol{x}$† - U‡ - U‡ - $\boldsymbol{d}$† （△）による場合，$\eta \geq 0.6$ に対する結果を見てみると，収束率は，γ が小さいときほぼ同じであるが，γ の増大とともに劣化しており，平均学習回数も全般的に悪くなっている．また，$\eta = 0.2$ に対する比較では，収束率は全般的に 10 % 程度低く，平均学習回数も悪いことが分かる．そして，片極型ネットワークによる結果（○）は，$\boldsymbol{x}$ - U - U‡ - $\boldsymbol{d}$† による結果（△）と同様の傾向にあるが，前者による収束率は γ が大きくなるにつれてより早めに劣化することが読みとれる．

3.7 結言

本章では，前章で述べた学習開始時における入力定義域分割の極性による違いが，荷重更新や解の実現にどのように影響するかを 2 − 2 − 1 ネットで XOR 問題を解かせる場合について調べ，片極型の場合，両極型に比べて良好な収束を与える初期値の範囲が狭くなる傾向にあることや，得られる解の種類が偏る恐れがあることを指摘した．また，ユニット特性がしきい値関数で与えられる場合について，XOR 問題の解領域を法ベクトルの集合で定式化して階層間の関係を明らかにし，解領域は極性によって大きく異なることを指摘した．この結果から，ユニット特性がシ

グモイド関数で与えられる通常の多層ネットで一般的な課題をBP学習させる場合，アトラクタや解領域はユニット極性に依存することが示唆される．

一方，片極型ネットと両極型ネットとで，荷重を(3.8)(3.10)式の関係で与えた場合，入力定義域での分離超平面による分割や活性領域の幅は等価になることを述べた．そして，この場合，フィードフォワード特性は同一となって，提示パターンに対する誤差特性が等しくなることを指摘した．これに基づいて，解領域の大きさはネットの極性によらず同一となることを指摘した．しかし，入力定義域の分割を幾何学的に等価にして学習を開始したとしても，荷重の更新特性が異なるため，収束挙動は一致しないことを明らかにした．つぎに，中間ユニットの極性が同じで，最終の出力ユニットのみの極性が異なる2つのネットワークに同じ初期荷重を与えて学習させた場合，収束挙動は一致することを証明した．そして，両極型ネットワークと片極型ネットワークのそれぞれに対して，誤差特性や収束挙動が同じになるネットワーク構成とそれへのパターン提示をとりまとめた．さらに，以上のことをシミュレーションにより検証した．また，シミュレーションにより，規模の大きなネットワークの初期値を通常のように乱数で与える場合，入力定義域の初期分割が幾何学的に等価でも，両極型ネットワークは，片極型ユニットをもとに構成されるネットワークに比べて，良好な収束を示す傾向にあることを述べた．それゆえ，大きな規模のネットワークが要請される問題を学習させる場合，両極型ネットワークよるアプローチは，後者に比べて，収束の点で有利と考えられる．

第4章　多層ネットの許容解濃度に関する考察

4.1　緒言

両極型ユニットとしては，一般に，出力範囲が -1 から 1 のものが使用されている．このときの入力定義域は $(-1,1)^m$ と一辺の長さが 2 の超立方体となって，その体積は 2^m と入力次元 m とともに増大する．一方，出力範囲が -0.5 から 0.5 の場合，入力定義域は $(-0.5,0.5)^m$ と単位超立方体となって，その体積は m の大きさに関係なく 1 となる．必要条件を満たす法ベクトルの集合は入力定義域を取り囲むように存在するから，入力定義域が拡大すれば，探索範囲が広がることになる．しかし，入力定義域の大きさの相違が多層ネットの許容解（全ての入力パターンに対してネット出力と教師との誤差がある許容範囲に入るときの荷重とバイアスの組で，実際にBP学習させて得られる収束解はこれに含まれる）の個数（濃度）や初期値の設定にどのように影響するかは定かでない．

また，ニューラルネットはハードウェア化されてその真価が発揮されることから，VLSI技術や光技術に基づいて，荷重やシグモイド関数，積和計算を実装したハードウェアが多数試作されている[19),23),55)]．この場合，シグモイド関数は，実装方式に依存して，固有の極性や値域幅をとる．そして，入力パターンは電気ないし光信号に変換されて提示されることになるが，許容される信号の大きさは回路的に制限される．同様な制限が荷重やバイアスの大きさについても存在する．したがって，計算機上で模擬的に構築された学習済みの多層ネットをハードウェア化する場合，実装されるシグモイド関数の特性との整合性を考慮しながら，入力パターン，荷重，バイアスをスケーリングする必要がある．さらに，学習可能なハードウェアの場合[7)]，これまで計算機シミュレーションを通して蓄積されてきた初期値や学習係数に関する経験値や知識をそこでの学習に活用することが重要となる．しかし，入力定義域が拡大縮小されたときのスケーリング法や

BP学習の収束挙動に関する議論は見当たらない.

以上の観点から，本章では，先ず，入力パターンがアフィン変換の関係にある学習課題に対して，層数や対応する層のユニット数が同じ多層ネットは，シグモイド関数の極性や値域幅の大きさによらず，等しい濃度の許容解をもつことを示す．次に，この結果を利用して，両極型ネットを対象に，入力パターンやシグモイド関数の値域が k 倍された場合，つまり入力定義域が k 倍に拡大された場合，バイアス荷重を同じにして，荷重を $1/k$ 倍すれば，分離能力や汎化能力を等価に保存できることを指摘する．さらに，荷重の初期値を上記のように設定した場合，バイアス荷重の更新に対する学習係数を $1/k^2$ 倍，また荷重更新に対する学習係数を $1/k^4$ 倍にすれば，収束挙動は等価になることを述べる．最後に，通常のように，荷重の初期値を同一分布幅の一様乱数で設定し，双方の更新に対する学習係数を等しく与えて学習させた場合，良好な収束を与える初期値は入力定義域の拡大とともに小さくなることをシミュレーションにより示す.

4.2　アフィン変換パターンに対する許容解の濃度

入力パターンと教師がそれぞれ $\boldsymbol{p}$ と $\boldsymbol{d}$ で与えられる訓練データの集合を $\{(\boldsymbol{p},\boldsymbol{d})\}$ のように表す．また，その課題に対して，シグモイド関数が $f(u)$ で与えられる多層ネットによる許容解の集合を $\Psi(\boldsymbol{p},\boldsymbol{d},f)$ と表記して解集合と呼ぶことにする．このとき，両極型および片極型ネットの解集合はそれぞれ $\Psi_B(\boldsymbol{p},\boldsymbol{d}_B,f_B)$ および $\Psi_U(\boldsymbol{q},\boldsymbol{d}_U,f_U)$ のように表される．ここに，教師 $\boldsymbol{d}_B$ の要素は -0.5 から 0.5，$\boldsymbol{d}_U$ の要素は 0 から 1 の範囲にあるものとする．この場合，既に報告したように，入力パターンと教師のそれぞれが，$\boldsymbol{q}=\boldsymbol{p}+\boldsymbol{1}/2$ および $\boldsymbol{d}_U=\boldsymbol{d}_B+\boldsymbol{1}/2$ と平行移動の関係にあれば，2つの解集合は $\Psi_B(\boldsymbol{p},\boldsymbol{d}_B,f_B)\Longleftrightarrow\Psi_U(\boldsymbol{q},\boldsymbol{d}_U,f_U)$，つまり1対1に対応して，2つのネットは同じ濃度の解集合をもつ[10].

上記は，訓練データが平行移動の関係にある課題は本質的に等価であり，この等価な学習課題に対して，同一構造の多層ネットは，シグモイド関数の値域幅が同じなら極性によらず，同じ濃度の解集合をもつことを意味している．これからの発展で，次のことが類推される．すなわち，入力パターンが $\{\boldsymbol{p}\}$ で与えられる課題と，入力パターンが $\{\boldsymbol{p}\}$ をアフィン変換して

$$\boldsymbol{q} \;=\; A\boldsymbol{p}+\boldsymbol{a} \tag{4.1}$$

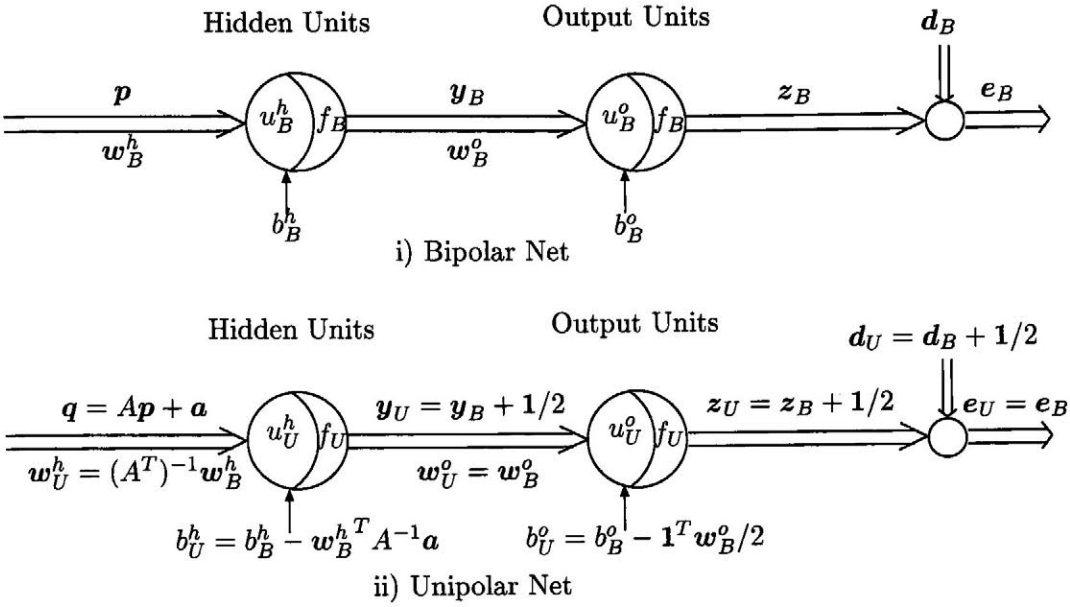


図 4.1: 前向き信号の等価な両極型と片極型ネットの概要図．($\boldsymbol{w}_B^h$ や $\boldsymbol{w}_B^o$ 等の表記における上付添字 h と o はそれぞれ中間および出力ユニットを指している．)

と与えられる課題は等価であり，これらの課題に対して，同一構造の多層ネットは，シグモイド関数の極性や値域幅によらず，同じ濃度の解集合をもつと考えられる．ここに，A は正則な変換行列，$\boldsymbol{a}$ は定数ベクトルである．以下では，このことを証明する．

4.2.1 ユニット極性が異なる場合

訓練データ $\{(\boldsymbol{p}, \boldsymbol{d}_B)\}$ に対する両極型ネットの解集合 $\Psi_B(\boldsymbol{p}, \boldsymbol{d}_B, f_B)$ と，入力パターンが (4.1) 式で与えられる $\{(\boldsymbol{q}, \boldsymbol{d}_U)\}$ に対する片極型ネットの解集合 $\Psi_U(\boldsymbol{q}, \boldsymbol{d}_U, f_U)$ を比較する．この場合，教師が $\boldsymbol{d}_U = \boldsymbol{d}_B + \boldsymbol{1}/2$ なる関係にあれば，$\Psi_B(\boldsymbol{p}, \boldsymbol{d}_B, f_B) \Longleftrightarrow \Psi_U(\boldsymbol{q}, \boldsymbol{d}_U, f_U)$ となって，双方のネットによる解集合は同じ濃度になることが以下のように示される．図 4.1 は，議論の流れを明らかにするため，3 層の場合を例に，前向き信号が等価な両極型ネットと片極型ネットの関係を概念的に示したものである．

先ず，双方で第 1 中間ユニットの荷重が

$$\boldsymbol{w}_U \;=\; (A^T)^{-1} \boldsymbol{w}_B \tag{4.2}$$

$$b_U \;=\; b_B - \boldsymbol{w}_B^T A^{-1}\boldsymbol{a} \tag{4.3}$$

の関係にあれば，総入力 u_U と u_B は $\boldsymbol{w}_U^T\boldsymbol{q} + b_U = \boldsymbol{w}_B^T\boldsymbol{p} + b_B$ のように等しくなることが導かれる．そして，この場合，$\boldsymbol{p}$ と $\boldsymbol{q}$ に対する第1中間ユニットの出力をそれぞれ y_B と y_U とすると，シグモイド関数の定義 (2.4)(2.3) 式から，

$$y_U \;=\; y_B + 1/2 \tag{4.4}$$

となって，片極型ユニットの出力は，両極型に比べて 1/2 大きくなることが分かる．

次に，第2中間層の場合，(4.4) 式より，y_B を要素とするベクトル $\boldsymbol{y}_B$ を両極型ユニットへの入力とすると，これを $\mathbf{1}/2$ 平行移動した $\boldsymbol{y}_U = \boldsymbol{y}_B + \mathbf{1}/2$ が片極型ユニットへ入力されることになる．この場合，(4.1) 式の変換行列と定数ベクトルがそれぞれ $A = I$（単位行列）と $\boldsymbol{a} = \mathbf{1}/2$ となることを考慮して，荷重を (4.2)(4.3) 式より $\boldsymbol{w}_U = \boldsymbol{w}_B, b_U = b_B - \mathbf{1}^T\boldsymbol{w}_B/2$ と与えれば，第2中間ユニットへの総入力は双方で等しいことが導かれる．このことは，分離超平面による入力定義域の分割が幾何学的に等価となるようにした (3.9) 式と一致する．同様のことが上位層ユニットについても繰返される．したがって，両極型ネットからの出力を $\boldsymbol{z}_B$ とすると，片極型ネットからの出力は $\boldsymbol{z}_U = \boldsymbol{z}_B + \mathbf{1}/2$ となる．

以上のことは，双方のネットにおける前向き信号の伝達特性が等価であることを示唆している．同時に，教師が $\boldsymbol{d}_U = \boldsymbol{d}_B + \mathbf{1}/2$ なる関係にある場合，$\{(\boldsymbol{p}, \boldsymbol{d}_B)\}$ に対する両極型ネットの誤差特性と，$\{(\boldsymbol{q}, \boldsymbol{d}_U)\}$ に対する片極型ネットの誤差特性は一致することを示している．したがって，双方のネットによる解集合は，$\Psi_B(\boldsymbol{p}, \boldsymbol{d}_B, f_B) \Longleftrightarrow \Psi_U(\boldsymbol{q}, \boldsymbol{d}_U, f_U)$ のように1対1に対応して，同じ濃度になることが分かる．

また，(4.4) 式の場合，(4.1) 式の関係にある $\boldsymbol{p}'$ と $\boldsymbol{q}'$ に対するユニット出力をそれぞれ y'_B と y'_U とすると，

$$y'_U - y_U \;=\; y'_B - y_B \tag{4.5}$$

となって，ユニット出力からみた入力パターンに対する分離能力は双方で等しくなる．さらに，$\boldsymbol{p}'$ と $\boldsymbol{q}'$ に対するネット出力をそれぞれ $\boldsymbol{z}'_B$ と $\boldsymbol{z}'_U$ とすると，$\boldsymbol{z}'_U - \boldsymbol{z}_U = \boldsymbol{z}'_B - \boldsymbol{z}_B$ となって，ネット出力からみた分離能力も等しくなることが分かる．言い換えると，2つのパターン集合 $\{\boldsymbol{p}\}$ と

$\{\boldsymbol{q}\}$ がアフィン変換の関係にあり，変換行列 A と定数ベクトル $\boldsymbol{a}$ が既知ならば，荷重を (4.2)(4.3) 式のように与えることにより，$\{\boldsymbol{p}\}$ に対する分離能力と $\{\boldsymbol{q}\}$ に対する分離能力を等価にする多層ネットが構築できる．ただし，この場合，双方のネットで，法ベクトルや活性領域幅は異なり，R^m における分離超平面や活性領域の配置も異なることに注意されたい．また，以下の場合についても，比較対象のネット間で，前向き信号の伝達特性や入力パターンに対する分離能力の等価性が上述と同様に導かれることに注意されたい．

4.2.2 ユニット極性が同一の場合

シグモイド関数が同じく (2.4) 式で与えられる 2 つの両極型ネットの解集合 $\Psi_B(\boldsymbol{p}, \boldsymbol{d}_B, f_B)$ と $\Psi_B(\boldsymbol{r}, \boldsymbol{d}_B, f_B)$ を比較する．ここに，$\boldsymbol{r}$ は $\boldsymbol{p}$ と (4.1) 式のようにアフィン変換の関係にあるものとする．この場合，第 1 中間層の荷重を (4.2)(4.3) 式のように与えれば，双方の総入力は同じになって，ユニット出力も同じになる．したがって，第 2 中間層以降の荷重が双方で等しければ，前向き信号の伝達特性は一致する．そして，教師も等しいから，双方の誤差特性は一致して，解集合 $\Psi_B(\boldsymbol{p}, \boldsymbol{d}_B, f_B)$ と $\Psi_B(\boldsymbol{r}, \boldsymbol{d}_B, f_B)$ は 1 対 1 に対応することになる．また片極型ネットについても，$\Psi_U(\boldsymbol{q}, \boldsymbol{d}_U, f_U)$ と $\Psi_U(\boldsymbol{s}, \boldsymbol{d}_U, f_U)$ の 1 対 1 対応が同様に導かれる．ここに，$\boldsymbol{s}$ は $\boldsymbol{q}$ とアフィン変換の関係にあるものとする．

4.2.3 ユニットの出力幅が異なる場合

シグモイド関数が k 倍されて $kf_B(\cdot)$ と与えられる両極型ネットの荷重をそれぞれ $\boldsymbol{w}_B^*$ と b_B^* で表す．そして，このネットに入力パターンが同じで教師のみが k 倍された訓練データ $\{(\boldsymbol{p}, k\boldsymbol{d}_B)\}$ を提示したときの解集合 $\Psi_B(\boldsymbol{p}, k\boldsymbol{d}_B, kf_B)$ と，$\Psi_B(\boldsymbol{p}, \boldsymbol{d}_B, f_B)$ を比較する．この場合，(4.1) 式の A と $\boldsymbol{a}$ はそれぞれ単位行列と零ベクトルになるから，(4.2)(4.3) 式より，第 1 中間ユニットが $\boldsymbol{w}_B^* = \boldsymbol{w}_B$, $b_B^* = b_B$ なる関係にあれば，その出力は $\boldsymbol{y}_B^* = k\boldsymbol{y}_B$ となる．第 2 中間層については，$A = kI$, $\boldsymbol{a} = \boldsymbol{0}$ を (4.2)(4.3) 式に代入して，$\boldsymbol{w}_B^* = \boldsymbol{w}_B/k$, $b_B^* = b_B$ となることから，総入力は双方で等しく，出力は k 倍されることが分かる．以下，同様な手順により，ネット出力が

$z_B^* = kz_B$ と k 倍されることが導かれる．したがって，誤差が k 倍されるという違いを許せば，$\Psi_B(\boldsymbol{p}, \boldsymbol{d}_B, f_B)$ と $\Psi_B(\boldsymbol{p}, k\boldsymbol{d}_B, kf_B)$ は1対1に対応することになる．同様に，片極型ネットについても，$\Psi_U(\boldsymbol{q}, \boldsymbol{d}_U, f_U)$ と $\Psi_U(\boldsymbol{q}, k\boldsymbol{d}_U, kf_U)$ の1対1対応が導かれる．

さらに，上述と4.2.2節の結果を両極型ネットの解集合 $\Psi_B(\boldsymbol{p}, \boldsymbol{d}_B, f_B)$ と $\Psi_B(\boldsymbol{r}, k\boldsymbol{d}_B, kf_B)$ に適用すれば，これらの要素も1対1に対応することが導かれる．また，片極型ネットについても，これと同様に，解集合 $\Psi_U(\boldsymbol{q}, \boldsymbol{d}_U, f_U)$ と $\Psi_U(\boldsymbol{s}, k\boldsymbol{d}_U, kf_U)$ の要素も1対1に対応することが導かれる．したがって，同一極性のネットは，構造が同じなら，アフィン変換された入力パターンに対し，シグモイド関数の値域幅に依らず，同じ濃度の解集合をもつことが分かる．

4.2.4　ユニットの極性と出力幅が異なる場合

シグモイド関数の極性や値域幅それに訓練データの観点から，以上の対応関係をまとめると図4.2のようになる．(図中の "$\Longleftrightarrow$" は1対1対応を示し，付記の数字はそれを証明した節を表す．)そして，図の対応関係から，例えば $\Psi_B(\boldsymbol{p}, \boldsymbol{d}_B, f_B) \Longleftrightarrow \Psi_U(\boldsymbol{s}, k\boldsymbol{d}_U, kf_U)$ が導かれて，ユニットの極性と出力幅がともに異なる場合でも解集合の濃度は等しくなることが分かる．

以上のことから，多層ネットは，同一構造，すなわち各層でのユニット個数が同じなら，シグモイド関数の極性や値域幅に関係なく，等価な学習課題に対して同じ濃度の解集合をもつことが結論づけられる．

4.3　スケーリングに関する考察

計算機上で学習済みの多層ネットをアナログ回路でハードウェア化する場合，実現可能なシグモイド特性に準拠して，入力パターンおよび荷重をスケーリングする必要がある．このスケーリングはすべての層の入力定義域にわたるから，各層の荷重の値はそれに応じて変換されなければならない．ここでは，先ず，このスケーリング問題に前述の結果を適用して，両極型ネットを対象に，入力パターンやシグモイド関数の値域幅が拡大縮小された場合でも，パターンの分離能力や汎化能力を等価に保存するネットが構築できることを述べる．

$$
\begin{array}{ccc}
\Psi_B(\boldsymbol{p}, k\boldsymbol{d}_B, kf_B) & \Longleftrightarrow & \Psi_B(\boldsymbol{r}, k\boldsymbol{d}_B, kf_B) \\
\Updownarrow \S 4.2.3 & \S 4.2.2 & \Updownarrow \S 4.2.3 \\
\Psi_B(\boldsymbol{p}, \boldsymbol{d}_B, f_B) & \Longleftrightarrow & \Psi_B(\boldsymbol{r}, \boldsymbol{d}_B, f_B) \\
\Updownarrow \S 4.2.1 & \S 4.2.2 & \Updownarrow \S 4.2.1 \\
\Psi_U(\boldsymbol{q}, \boldsymbol{d}_U, f_U) & \Longleftrightarrow & \Psi_U(\boldsymbol{s}, \boldsymbol{d}_U, f_U) \\
\Updownarrow \S 4.2.3 & \S 4.2.2 & \Updownarrow \S 4.2.3 \\
\Psi_U(\boldsymbol{q}, k\boldsymbol{d}_U, kf_U) & \Longleftrightarrow & \Psi_U(\boldsymbol{s}, k\boldsymbol{d}_U, kf_U) \\
 & \S 4.2.2 &
\end{array}
$$

図 4.2: 解集合間の 1 対 1 対応

また，BP 学習の収束能力は，学習係数や慣性係数の値に影響されるが，適切な値は学習課題に依存するため，その設定は経験に委ねざるを得ない現状にある．そのため，良好な学習過程をハードウェアで実現するには，これまでソフトウェア的に蓄積されてきた学習係数や慣性係数に関する経験値を如何に活用するかが重要となる．したがって，次に，入力パターンやシグモイド特性が拡大縮小されたときの BP 学習について，学習係数や慣性係数をどのように与えれば，収束挙動が等価になるかを議論する．

以上の 2 点を議論するにあたり，定義や仮定を次のように設ける．すなわち，計算機上で，(2.4) 式の $f_B(\cdot)$ をシグモイド関数として，訓練データ $\{(\boldsymbol{p}, \boldsymbol{d}_B)\}$ により既に BP 学習済みの両極型ネットを $\mathcal{N}$，またこれをハードウェアとして実現すべきネットを $\mathcal{N}^*$ と表す．さらに，$\mathcal{N}^*$ は $\mathcal{N}$ と同一構造として，$\mathcal{N}^*$ のシグモイド特性と訓練データをそれぞれ

$$f_B^*(u^*) = kf_B(u) \tag{4.6}$$

$$\{(\boldsymbol{p}^*, \boldsymbol{d}_B^*)\} = \{(k'\boldsymbol{p}, k\boldsymbol{d}_B)\} \tag{4.7}$$

のように定義する．この場合，k' と k は同じとしても議論の本質は失われない．したがって，以下では，便宜のため，$k' = k$ と仮定し，k を拡大率と呼ぶことにする．このとき，$\mathcal{N}^*$ の入力定義域は，$\mathcal{N}$ に比べて，すべての層で k 倍に拡大されることになる．

4.3.1　分離および汎化能力の等価なネット

入力パターン $\boldsymbol{p}^*$ と $\boldsymbol{p}$ は (4.1) 式において $A = kI, \boldsymbol{a} = \boldsymbol{0}$ の関係にある．したがって，$\mathcal{N}$ の荷重をそれぞれ $\boldsymbol{w}$ と b とするとき，(4.2)(4.3) 式に準じて，$\mathcal{N}^*$ の結合荷重 $\boldsymbol{w}^*$ とバイアス荷重 b^* をそれぞれ

$$\boldsymbol{w}^* = \boldsymbol{w}/k \tag{4.8}$$

$$b^* = b \tag{4.9}$$

と与えれば，

$$\boldsymbol{y}^* = k\boldsymbol{y} \tag{4.10}$$

となることが導かれる．ここに，$\boldsymbol{y}$ と $\boldsymbol{y}^*$ はそれぞれ $\mathcal{N}$ と $\mathcal{N}^*$ の対応する層のユニット出力である．したがって，$\mathcal{N}^*$ の各層に対する入力は，$\mathcal{N}$ の対応する層の入力に比べて，k 倍されることになる．これは，2つのネットの前向き信号伝達特性が等価であることを意味している．

一方，荷重が (4.8)(4.9) 式で与えられるとき，(2.17)(2.18) 式から，法ベクトルや活性領域幅も

$$\boldsymbol{h}^* = k\boldsymbol{h} \tag{4.11}$$

$$g^* = kg \tag{4.12}$$

と k 倍されることが分かる．このことは，双方の入力定義域が法ベクトルや活性領域により幾何学的に相似な形で分割されることを意味する，と同時に入力パターンに対する分離能力が等価となることを意味している．

そして，$\mathcal{N}$ と $\mathcal{N}^*$ において，ネット出力と教師の誤差をそれぞれ e および e^* と記すとき，

$$e^* = ke \tag{4.13}$$

となって，両者による誤差特性は等価になることが導かれる．このことは，入力パターンが k 倍された場合，教師も k 倍して，荷重を (4.8)(4.9) 式のように与えれば，$\mathcal{N}^*$ と $\mathcal{N}$ の汎化能力が等価になることを示唆している．

4.3.2 等価な収束挙動を与える学習係数

入力定義域が (4.6)(4.7) 式のように拡大されるとき，荷重の初期値を (4.8)(4.9) 式の関係で与え，訓練データを $\mathcal{N}$ と $\mathcal{N}^*$ に同一順序で提示して学習させる場合を考える．そして，$\mathcal{N}^*$ の学習係数や慣性係数を $\mathcal{N}$ のものとどのような関係で与えれば，両者の収束挙動が等価になるかを考察する．

BP 学習において，$\mathcal{N}$ の出力ユニットの結合荷重 $\boldsymbol{w}$ とバイアス荷重 b は，逆伝搬誤差 $\delta = f'_B(\cdot)e$ をもとに，

$$\Delta \boldsymbol{w} = \eta\delta\boldsymbol{x} + \alpha\Delta\boldsymbol{w}(old) \tag{4.14}$$

$$\Delta b = \mu\delta + \alpha\Delta b(old) \tag{4.15}$$

と更新される．ここに，$f'_B(\cdot)$ は $f_B(\cdot)$ の微分，η と μ はそれぞれ結合荷重とバイアス荷重更新の学習係数，α は慣性係数，$\boldsymbol{x}$ は出力ユニットへの入力ベクトルである．また，$\mathcal{N}^*$ の出力ユニットの結合荷重 $\boldsymbol{w}^*$ とバイアス荷重 b^* は，逆伝搬誤差 $\delta^* = f'^*_B(\cdot)e^*$ をもとに，

$$\Delta \boldsymbol{w}^* = \eta^*\delta^*\boldsymbol{x}^* + \alpha^*\Delta\boldsymbol{w}^*(old) \tag{4.16}$$

$$\Delta b^* = \mu^*\delta^* + \alpha^*\Delta b^*(old) \tag{4.17}$$

と更新される．ここに，η^*，μ^*，α^*，$\boldsymbol{x}^*$ は $\mathcal{N}^*$ に対して上記と同様に定義される量である．

結合荷重とバイアス荷重の初期値をそれぞれ (4.8)(4.9) 式の関係で与えた場合，前節で述べたように，$\mathcal{N}$ と $\mathcal{N}^*$ の前向き信号伝達は (4.10) 式のように等価となり，誤差特性も (4.13) 式のように等価になる．したがって，(4.13) 式と (4.6) 式から，$\mathcal{N}$ と $\mathcal{N}^*$ の逆伝搬誤差について，

$$\delta^* = k^2\delta \tag{4.18}$$

なる関係が得られる．この関係と (4.10) 式を (4.14) 式から (4.17) 式に適用すると，最初の提示パターンに対して，$\Delta\boldsymbol{w}(old)$, $\Delta b(old)$, $\Delta\boldsymbol{w}^*(old)$, $\Delta b^*(old)$ はゼロとなることから，次の関係が導かれる．すなわち，結合荷重とバイアス荷重更新の学習係数をそれぞれ

$$\eta^* = \eta/k^4 \tag{4.19}$$

$$\mu^* = \mu/k^2 \tag{4.20}$$

と設定すれば，荷重の変化分について，

$$\Delta \boldsymbol{w}^* = \Delta \boldsymbol{w}/k \tag{4.21}$$

$$\Delta b^* = \Delta b \tag{4.22}$$

なる関係が得られる．(4.21)(4.22) 式は，$\mathcal{N}$ と $\mathcal{N}^*$ の出力ユニットの荷重が，更新後も，更新前の (4.8)(4.9) 式と同じ関係を保存していることを意味している．同様に，中間ユニットについても，更新後の荷重が (4.8)(4.9) 式の関係を保存することが導かれる．したがって，最初の提示パターンによる更新後も，$\mathcal{N}$ と $\mathcal{N}^*$ の前向き信号伝達特性は等価となり，(4.10) 式が保証されることになる．

2番目の提示パターンについては，$\Delta \boldsymbol{w}^*(old) = \Delta \boldsymbol{w}(old)/k$, $\Delta b^*(old) = \Delta b(old)$ なる関係のもとで，(4.14) 式から (4.17) 式のように更新される．この場合，慣性係数を $\mathcal{N}$ と $\mathcal{N}^*$ の間で

$$\alpha^* = \alpha \tag{4.23}$$

と等しく設定すれば，出力ユニットの荷重の変化分について，$\Delta \boldsymbol{w}^* = \eta^* \delta^* \boldsymbol{x}^* + \alpha^* \Delta \boldsymbol{w}^*(old) = (\eta \delta \boldsymbol{x} + \alpha \Delta \boldsymbol{x}(old))/k = \Delta \boldsymbol{w}/k$ のように (4.21) 式が成立し，バイアス荷重の変化分についても (4.22) 式が成立する．同様に，中間ユニットの荷重の変化分についても (4.21)(4.22) 式が成立することが導かれる．

上述の議論を以降の提示パターンについても繰返すことにより，次の結論に達する．すなわち，$\mathcal{N}$ と $\mathcal{N}^*$ において，荷重の初期値をそれぞれ $\boldsymbol{w}^* = \boldsymbol{w}/k$ と $b^* = b$ の関係で与えて，学習係数を $\eta^* = \eta/k^4$ および $\mu^* = \mu/k^2$，また慣性係数を $\alpha^* = \alpha$ と設定した場合，2つのネットによる BP 学習の収束挙動は等価になる．

4.3.3　バイアス荷重の駆動入力を考慮する場合

通常，ユニットへの総入力は (2.1) 式のように表現される．これはバイアス荷重の駆動入力を 1 なる固定値とみなすことに他ならない．一方，ハードウェア化においては，結合荷重とバイアス荷重は区別されることなく同一方式で実現されることが多い[7),19),23),55)]．この場合，(2.1) 式を $\boldsymbol{w}^T \boldsymbol{p} + b = (\boldsymbol{w}^T/k)(k\boldsymbol{p}) + (b/k)k$ と変形して分かるように，バイアス荷重の駆動入力を k なる

固定入力として扱う必要がある．その際，分離能力や汎化能力を等価にする荷重は，$\boldsymbol{w}^* = \boldsymbol{w}/k$，$b^* = b/k$ と同じく $1/k$ 倍にスケーリングすればよいことが導かれる．また，等価な収束挙動については，結合荷重およびバイアス荷重更新の学習係数を $\eta^* = \eta/k^4$ および $\mu^* = \mu/k^4$ と同じく $1/k^4$ 倍すればよいことが導かれる．

4.4 スケーリングと適切な初期値

前節では，入力定義域がスケーリングされた場合，初期値を (4.8)(4.9) 式のように結合荷重とバイアス荷重で区別して設定し，更新アルゴリズムについても，学習係数を結合荷重とバイアス荷重で (4.19)(4.20) 式のように区別して与えれば，収束挙動は等価になることを述べた．しかし，現状では初期値や学習係数の設定法が確立されていないため，通常，初期値は結合荷重とバイアス荷重とで区別されることなく同じ分布幅の一様乱数で与えられ，学習係数についても同じ値が用いられる．

このように通常の形態で BP 学習させる場合，良好な収束を与える初期値の大きさには適切な範囲が存在することが知られている[40),49)]．しかし，その範囲は学習課題に依存して特定が容易でないため，一般に，初期値としては小さな乱数を採用することが多い．これに関して，入力定義域が拡大された場合，前節までの議論から，少なくとも荷重の初期値はさらに小さい方が収束は良好になると考えられるが，詳細は定かでない．したがって，ここでは，通常のように，荷重の初期値を同一分布幅 $(-\gamma, \gamma)$ の一様乱数で設定し，結合荷重とバイアス荷重更新の学習係数を等しく与えて学習させた場合，良好な収束を与える初期値の大きさ γ が入力定義域のスケーリングにどのように影響されるかをシミュレーションにより調べる．

シミュレーションは，8 次のパリティ問題を対象に，$8-30-1$ の両極型ネットで，学習係数を $\eta = 0.5$，慣性係数を $\alpha = 0.9$ として，$\gamma = 0.3, 0.5, 1.0, 1.5, 2.0, \cdots, 5.0$ の場合について行った．その際，教師は $\pm 0.4k$ として与え，収束判定基準は 256 個の入力パターンの全てに対して誤差の絶対値が $0.1k$ 未満に収まることとした．図 4.3 は，拡大率 k を 1.0 （○），3.0 （△），5.0 （□），10.0 （×） 倍としたときの収束率 (Convergence Rate : 50 回の試行における収束回数の比) と収束したときの平均学習回数 (Epoch) の結果である．図より，$k = 3$ のとき $1 \leq \gamma \leq 2$，また

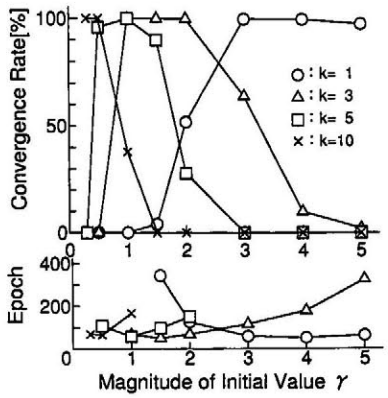


図 4.3: 初期値設定とスケーリングからみた典型的な BP 学習の収束

$k = 5$ のとき $0.5 \leq \gamma \leq 1.5$ なる範囲の初期値に対して良好な収束が得られていることが読みとれる．また，その他の拡大率についても，k が大きくなるにつれて，良好な収束を与える γ は小さくなっていることが分かる．これは，η や α の値を変えたり，学習課題を変えた場合でも同様であった．以上は，入力定義域が拡大したときの適切な初期値に対する設定指針を示唆している．

4.5　結言

本章では，入力パターンがアフィン変換の関係にあり本質的に等価と考えられる学習課題に対して，同一構造の多層ネットは，シグモイド関数の極性や値域幅の大きさによらず，等しい濃度の許容解をもつことを指摘した．また，この結果を利用して，両極型ネットを対象に，入力定義域が k 倍にスケーリングされた場合，バイアス結合を同じにして，荷重を $1/k$ 倍すれば，分離能力や汎化能力は等価になることを指摘した．さらに，荷重を上記のように初期化した場合，バイアス荷重更新に対する学習係数を $1/k^2$ 倍，また荷重更新に対する学習係数を $1/k^4$ 倍にすれば，収束挙動は等価になることを述べた．最後に，通常のように，荷重の初期値を同一分布幅の一様乱数で設定し，双方の更新に対する学習係数を等しく与えて学習させた場合，良好な収束を与える初期値は入力定義域の拡大とともに小さくなることをシミュレーションにより示した．そして，これらの結果が多層ネットのハードウェア化に際して有用になることを述べた．

第5章　両極型ネットの優位性

5.1　緒言

パターン分類問題において多層ネットによる学習が収束して解となる荷重が得られた場合，本質的にいくつかのユニットについては，その荷重の形成する分離超平面が入力集合を通って個々の入力を有意に分離していると考えられる．したがって，入力集合を有意に分離する分離超平面の分布領域が得られれば，BP 法の収束問題を初期値やアトラクタの観点から掘り下げて議論できると思われる．しかし，この分布領域を厳密に求めるのは困難である[51]．

そこで，本章では，入力集合の代わりにこれに外接する（超）直方体で近似し，その内部を分離超平面が通るための条件を分離条件と呼ぶことにして，これを法ベクトルの集合により定式化する．そして，この分離条件を満たす法ベクトルの分布領域を 2 次元の場合について図示する．この図から，高次元ベクトル空間における分離条件の幾何学的な形状に関する知見が得られ，入力パターン集合が単に平行移動したり，ネット極性が違うだけでも，分離条件は大きく異なることがわかる．また，荷重の初期値を通常のように平均がゼロの一様乱数で与えたときの法ベクトルの分布領域についても図示する．以上の二つの分布領域を同一のベクトル空間で比較することにより，通常の初期値設定による BP 学習の収束は，片極型ネットに比べて，両極型ネットの方が優位となることが導かれる．さらに，双方のネットで入力集合の分割が等価となるように初期値設定した場合でも，片極型ネットに比べて，両極型ネットによる収束は幅広い学習係数に対して良好となることをシミュレーションにより示し，分離条件を考慮した初期値設定の効果についても考察する．

5.2　幾何学的観点からみた学習の収束

多層ネットには，一般に，訓練パターンの規定する入出力関係を満たす解（$\boldsymbol{w}$ や b の組合せ）は多数存在する．これは，法ベクトルの解（$\boldsymbol{h}$ の組合せ）についても同様である．一方，BP 法は最急降下法に基づく多峰的な誤差曲面の最小値探索法であり，その収束能力は荷重の初期値，等価的に法ベクトルの初期配置や初期活性領域の広さに依存する．したがって，BP 法の場合，初期値をこれらの解のアトラクタ領域に高い確率で設定できれば，学習の収束する割合は高くなると考えられる．

パターン分類問題を多層ネットで学習させた場合，前述のように，ユニットのなす分離超平面は入力集合を分離している考えられる．しかし，ユニットが冗長にある多層ネットについては，解を持つ場合でも，ユニットによっては分離超平面が入力集合を通らない可能性がでてくる．これは，冗長な層を持つ多層ネットについても同様である．たとえば，AND や OR 等の線形分離可能な問題は本質的に 2 層のネットで解けるが，これを 3 層ネットで学習させた場合，第 4 章でのアフィン変換に関する議論からわかるように，冗長な層のユニットについては，必ずしも分離超平面が入力集合を通る必要はない．そのため，以下では，このような分離超平面は解の本質に深く関与しないとして除外して考える．

また，入力集合を層毎に見てみると，第 1 中間層への入力集合は，多層ネットへの提示入力パターン集合そのものであり，学習期間中変わることがないのに対して，第 2 中間以降の層への入力集合は，前層のユニットでの荷重更新により学習期間中たえず変化する．そのため，第 1 中間層については，提示入力パターン集合を最終的に定まる入力集合として取り扱えるが，第 2 中間以降の層については，最終的に確定する入力集合をあらかじめ知ることはできない．

以降では，便宜のため，入力，中間，出力層からなる 3 層ネットを考え，中間層と出力層に対する分離条件を議論する．しかし，そこで展開される議論は，一般性を失うことなく，3 層以上の多層ネットに対しても適用できる．この場合，第 2 中間以降の層に対する議論は，3 層ネットの出力層に対する議論と等価となる．

5.3 中間層における分離条件

ここでは，提示入力パターン集合を外接直方体で近似し，その内部を分離超平面が通るための分離条件を求める．そして，平行移動の関係にある三つの入力集合を近似する外接直方体に対する分離条件を2次元の場合を例に図示する．その結果，分離条件を満たす法ベクトルの分布は，入力集合が単に平行移動されただけでも大きく異なることを指摘する．

5.3.1 入力集合の直方体近似

多層ネットの訓練パターンとして，S 組の提示入力パターンと教師の対を考え，$s\,(=1,2,\cdots,S)$ 番目の入力パターンを $\boldsymbol{x}_s=[x_{s1},x_{s2},\cdots,x_{sm}]^{\mathrm{T}}$，入力パターン集合を $\mathcal{X}=\{\boldsymbol{x}_s\}$ と表す．

このとき，入力パターンの第 i 成分の最大値と最小値を

$$x_i^{max}=\max_s x_{si},\quad x_i^{min}=\min_s x_{si} \tag{5.1}$$

として，中心 $\boldsymbol{c}=[c_1,c_2,\cdots,c_m]$ と頂点 $\boldsymbol{p}_k=[p_{k1},p_{k2},\cdots,p_{km}]$ がそれぞれ

$$c_i = \frac{(x_i^{max}+x_i^{min})}{2} \quad i=1,2,\cdots m \tag{5.2}$$

$$p_{ki} = x_i^{max}\ or\ x_i^{min} \quad k=1,2,\cdots,2^m \tag{5.3}$$

と定まる $\mathcal{X}$ の外接（超）直方体を Π と定義する．また，$\mathcal{X}$ を $\boldsymbol{x}_s^*=\boldsymbol{x}_s-\boldsymbol{c}$ のように平行移動したものを $\mathcal{X}^*=\{\boldsymbol{x}_s^*\}$，$\Pi$ を平行移動して，頂点が $\boldsymbol{p}_k^*=\boldsymbol{p}_k-\boldsymbol{c}$ で与えられる直方体を Π^* と定義する．この Π^* は $\mathcal{X}^*$ の外接直方体で，その中心は R^m の原点 O に一致する．

さらに，Π^* は，原点 O と頂点 $\boldsymbol{p}_k^*$ を対頂点とする 2^m 個の直方体 Γ_k^* に分けられる．このとき，頂点 $\boldsymbol{p}_k^*$ を通る分離超平面に随伴する法ベクトルの軌跡は，原点とその頂点を結ぶ線分を直径とする球面となるため，Γ_k^* の内部を通る分離超平面 H_k^* の法ベクトル $\boldsymbol{h}_k^*$ は，

$$\Phi_k^*=\{\boldsymbol{h}_k^* \mid \|\boldsymbol{h}_k^*-\frac{\boldsymbol{p}_k^*}{2}\|\le\|\frac{\boldsymbol{p}_k^*}{2}\|\} \tag{5.4}$$

と導かれる．したがって，直方体 Π^* の内部を通る分離超平面 H^* の法ベクトル $\boldsymbol{h}^*$ は，Φ_k^* の和集合の要素として，次のように与えられることになる．

$$\boldsymbol{h}^* \in \Omega^*=\bigcup_{k=1}^{2^m}\Phi_k^* \tag{5.5}$$

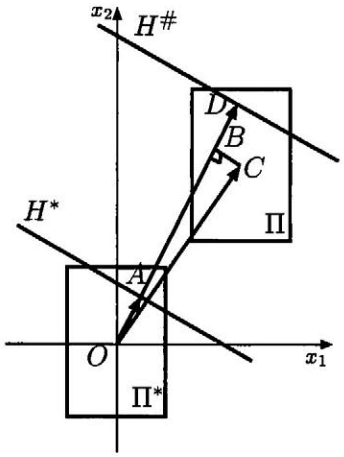


図 5.1: 分離超平面による幾何学的に等価な入力定義域の分割

5.3.2　中間ユニットに対する分離条件

直方体 Π^* と Π が図5.1のように，それぞれ分離超平面 H^* と $H^\sharp$ により幾何学的に等価な形で分割される場合を考える．ここに，$\overrightarrow{\mathrm{OD}}$ と $\overrightarrow{\mathrm{OA}}$ はそれぞれ $H^\sharp$ と H^* に随伴する法ベクトル $\boldsymbol{h}^\sharp$ と $\boldsymbol{h}^*$ で，C は直方体 Π の中心，B は線分 OD と中心 C からそれへの垂線との交点である．この場合，分割が幾何学的に等価であることから，$\overrightarrow{\mathrm{OA}} = \overrightarrow{\mathrm{BD}}$ となって，$\boldsymbol{h}^\sharp = \overrightarrow{\mathrm{OB}} + \boldsymbol{h}^*$ が成り立つ．また，線分 OB は線分 OC の射影であるから，

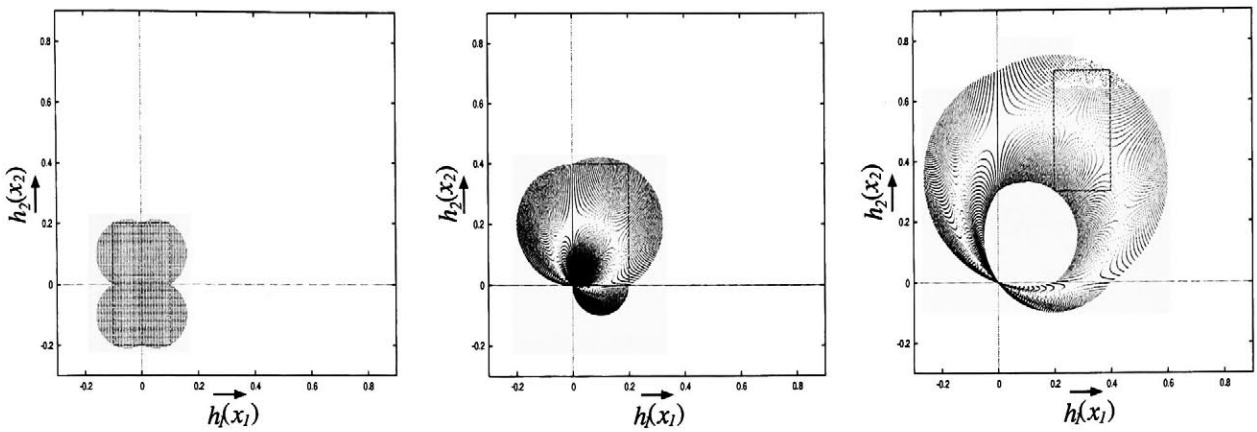
$$\boldsymbol{h}^\sharp \;=\; (1 + \frac{\boldsymbol{c}^{\mathrm{T}}\boldsymbol{h}^*}{\|\boldsymbol{h}^*\|^2})\boldsymbol{h}^* \tag{5.6}$$

となる．したがって，直方体 Π を通る分離超平面 $H^\sharp$ の法ベクトル $\boldsymbol{h}^\sharp$ は，Φ_k^* の (5.6) 式による写像 $\Phi_k^\sharp$ の和集合 $\Omega^\sharp = \bigcup \Phi_k^\sharp$，すなわち，

$$\Omega^\sharp = \{\boldsymbol{h}^\sharp \mid \boldsymbol{h}' \in \Omega^*,\ \boldsymbol{h}' = (1 - \frac{\boldsymbol{c}^{\mathrm{T}}\boldsymbol{h}^\sharp}{\|\boldsymbol{h}^\sharp\|^2})\boldsymbol{h}^\sharp\} \tag{5.7}$$

の要素で与えられる．以上のことから，中間ユニットに対する分離条件は，(5.7) 式の法ベクトル集合 $\Omega^\sharp$ により定式化されることがわかる．

以下では，平行移動の関係にある三つの入力パターン集合 $\mathcal{X}_{\mathrm{org}}$, $\mathcal{X}_{\mathrm{vtx}}$, $\mathcal{X}_{\mathrm{off}}$ を考え，それぞれを近似する合同な外接直方体 Π_{org}, Π_{vtx}, Π_{off} が次のように定義される場合について，(5.7) 式の分離条件を満たす法ベクトル集合の幾何学的な分布 $\Omega^\sharp$ の相違を調べる．すなわち，(a) 外接直方体



(a) $\Omega^\sharp_{\rm org}$ (b) $\Omega^\sharp_{\rm vtx}$ (c) $\Omega^\sharp_{\rm off}$

図 5.2: 分離条件を満たす法ベクトルの分布

$\Pi_{\rm org}$ の中心が原点にある場合 ($\boldsymbol{c}=\boldsymbol{0}$) の分布 $\Omega^\sharp_{\rm org}$, (b) $\Pi_{\rm vtx}$ の頂点が原点に一致する場合 ($\boldsymbol{c}=\boldsymbol{p}^*_1$) の分布 $\Omega^\sharp_{\rm vtx}$, (c) $\Pi_{\rm off}$ が原点を含まない場合 ($\|\boldsymbol{c}\| \geq \|\boldsymbol{p}^*_k\|/2$) の分布 $\Omega^\sharp_{\rm off}$ の違いを調べる.

図 5.2 は, $\Pi_{\rm org}$, $\Pi_{\rm vtx}$, $\Pi_{\rm off}$ がいずれも縦 0.4, 横 0.2 の長方形で, その中心 $\boldsymbol{c}$ がそれぞれ (a) $[0,0]^{\rm T}$, (b) $[0.1,0.2]^{\rm T}$, (c) $[0.3,0.5]^{\rm T}$ のときの法ベクトル ($m=2$ 次元) の分布例である. 具体的には, 第 1 成分と第 2 成分をそれぞれ 0.03 刻みで動かし, (5.5) 式を満たす点 (法ベクトル) $\boldsymbol{h}^*$ を求めて描いたものが図 5.2(a) で, さらにこれを (5.6) 式で写像して $\boldsymbol{h}^\sharp \in \Omega^\sharp$ と求めたのが図 5.2(b) と (c) である. このときの図 5.2(b) と (c) における濃淡は (5.6) 式の非線形性による.

図 5.2 の結果から, 単に平行移動しただけの合同な直方体を等価に分割するだけでも, その分離超平面に随伴する法ベクトルの分布領域は大きく異なることがわかる. とくに, 原点を含まない $\Pi_{\rm off}$ に対する $\Omega^\sharp_{\rm off}$ については, 図 5.2(c) に示すように, 環状に分布して, 空洞が必ず存在し, 原点付近における分布領域は極めて狭くなっている. これは, 他の二つの分布に比べて, 際だって異なる特徴となっている. そして, 直方体が原点から遠ざかるにつれて, 空洞が広がるため, 環は大きくなって, 全体的に細くなる.

以上のように, 入力集合が単に平行移動しただけでも, 分離条件は大きく異なる. そして, このことは, 入力パターン集合が同じ位相構造の等価な学習課題でも, 分離条件が入力パターン集合の R^m における幾何学的な広さ, 方向, 位置に依存して変わることを示唆している.

5.4　出力層における広義の分離条件

中間層への入力集合は学習期間中変わることはないが，出力層への入力集合は，学習期間中，更新される中間ユニットの荷重に依存して推移する．そのため，出力ユニットでは，解が得られるときの入力集合は収束するまで確定できない．しかし，収束して最終的に定まる入力集合は必ず後出の入力定義域に含まれる．

したがって，ここでは，出力ユニットに対する分離条件を，その分離超平面が入力定義域を通るときの法ベクトル集合として求める．この法ベクトルの分布は，分離超平面が最終的に確定した入力集合の外接直方体を通るときの法ベクトルの分布よりも，かなり広い．その意味で，ここで得られる法ベクトル集合については，広義の分離条件とよんで，中間層に対する分離条件と区別する．

5.4.1　両極型ネットに対する広義の分離条件

両極型ネットの場合，出力ユニットに対する入力定義域は，シグモイド関数の定義 (2.4) 式より，$X_B = (-0.5, 0.5)^m$ のように原点を中心とする一辺の長さが 1 の単位超立方体となる．このとき，(5.4) 式と同様に，原点と頂点を結ぶ線分を直径とする超球 Φ_{Bk} が定義できて，この和集合 $\Omega_B = \bigcup \Phi_{Bk}$ が X_B を通る分離超平面に随伴する法ベクトルの集合となることが導かれる．この集合は，さらに，簡略化されて，

$$\Omega_B \;=\; \{\boldsymbol{h}_B \,|\, 2\sum_{i=1}^{m}(h_{Bi})^2 \le \sum_{i=1}^{m}|h_{Bi}|\} \tag{5.8}$$

と表現される[48]．ここに，h_{Bi} は $\boldsymbol{h}_B$ の第 i 番目の成分である．したがって，(5.8) 式で定義される Ω_B が，両極型ネットの出力ユニットに対する広義の分離条件となる．

5.4.2　片極型ネットに対する広義の分離条件

片極型ネットの出力ユニットに対する入力定義域は，シグモイド関数の定義 (2.3) 式より，$X_U = (0, 1)^m$ となって，X_B を $\boldsymbol{c} = \mathbf{1}/2$ だけ平行移動した単位超立方体となる．ここに，$\mathbf{1}$ は $[1, 1, \cdots, 1]^{\mathrm{T}}$

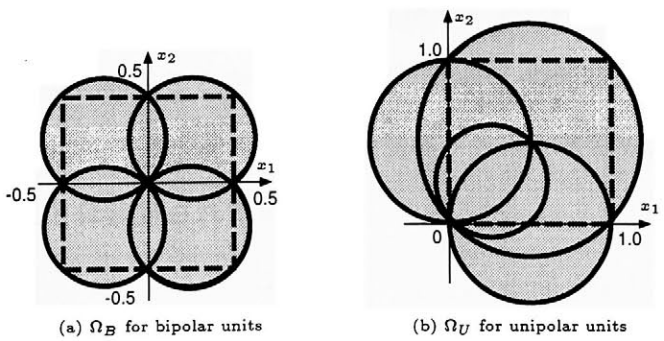
のように要素がすべて 1 の m 次元ベクトルである．したがって，片極型ネットの出力ユニットに対する広義の分離条件 Ω_U は，$\boldsymbol{c} = \mathbf{1}/2$ としたときの (5.6) 式による Φ_{Bk} の写像 Φ_{Uk} の和集合 $\bigcup\Phi_{Uk}$，もしくは，これを簡略化した表現で次のように与えられる．

$$\begin{aligned}\Omega_U &= \{\boldsymbol{h}_U \,|\, 2\sum_{i=1}^{m}(h'_{Ui})^2 \leq \sum_{i=1}^{m}|h'_{Ui}|, \\ &\quad \boldsymbol{h}'_U = (1 - \frac{\sum_{i=1}^{m} h_{Ui}}{2\|\boldsymbol{h}_U\|^2})\boldsymbol{h}_U\}\end{aligned} \tag{5.9}$$

5.4.3 広義の分離条件の極性による相違

広義の分離条件の極性による相違を明確にするため，$m = 2$ 次元のときの Ω_B や Ω_U を図示すると，それぞれ図 5.3 の (a) と (b) の点表示部のようになる．ここに，図中の円は Φ_{Bk} や Φ_{Uk} $(k = 1, 2, 3, 4)$ を意味している．図より，法ベクトルの集合 Ω_B と Ω_U はそれぞれ入力定義域 X_B と X_U を取り囲むように存在するが，両者の存在領域は異なり，$\boldsymbol{h}_B \in \Omega_B$ は各象限で均等に存在し任意の方向をとれるのに対して，$\boldsymbol{h}_U \in \Omega_U$ は大部分が第 1 象限に偏在し，第 3 象限には存在しないことがわかる．このことと (2.17) 式より，両極型の場合，荷重の符号として任意の組合せが可能であるのに対して，片極型の場合，これらのすべてが同一符号となることはなく，符号の組合せが制約されることが導かれる．このことは Watanabe らの実験結果[33]からも裏付けられる．

また，両極型の場合，$\boldsymbol{h}_B \in \Omega_B$ の方向と長さの関係は任意の象限で等しいのに対して，片極型



(a) Ω_B for bipolar units　　(b) Ω_U for unipolar units

図 5.3: 広義の分離条件の極性による分割の違い

の場合，$\boldsymbol{h}_U \in \Omega_U$ の方向と長さは，どの象限に存在するかで大きく異なる．とくに，入力定義域の中心を通る分離超平面については，両極型の場合，$\boldsymbol{h}_B = \mathbf{0}$ の法ベクトルで与えられるが，片極型の場合，$\mathbf{1}/4$ を中心とする半径 $\|\mathbf{1}/4\|$ の超球面上の法ベクトルで与えられるという違いがある[10]．

5.5　学習収束における両極型ネットの優位性

片極型と両極型ネットの双方に同じ入力パターンを提示した場合，中間ユニットに対する分離条件は法ベクトル集合 $\Omega^{\natural}$ で与えられ，これは極性に依存しない．しかし，出力ユニットに対する広義の分離条件は，極性により Ω_U もしくは Ω_B と異なる．

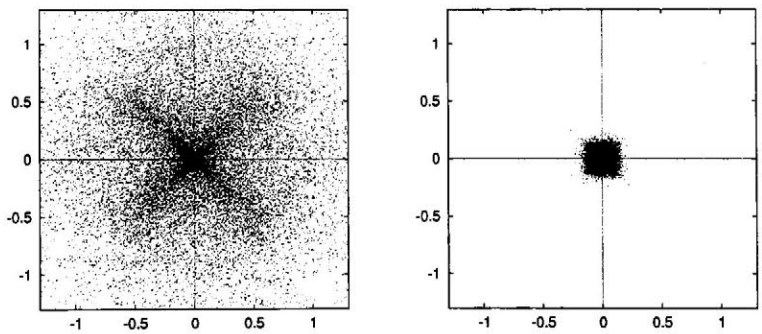
通常，入力パターンの提示については，使用するネットの極性に応じて暗黙の正規化が行われている．たとえば，XOR 問題を学習させる場合，片極型ネットに対しては，入力パターンを $\mathcal{X}_U^{\mathrm{xor}}$ = { (0,0), (0,1), (1,0), (1,1) }，また両極型ネットに対しては $\mathcal{X}_B^{\mathrm{xor}}$ = { $(-1,-1)$, $(-1,1)$, $(1,-1)$, $(1,1)$ } のように提示することが慣習的に行われている．このように，ネット極性に応じて正規化した入力パターンを提示した場合，中間ユニットに対する分離条件も極性に依存する．この場合，すべての層で，(広義の) 分離条件は極性によって異なることになる．また，分離条件の広さは，入力集合の大きさに依存する．

そこで，以下では，入力パターンはあらかじめ正規化されており，その外接直方体は，両極型のとき $[-1/2, 1/2]^m$，片極型のとき $[0,1]^m$ と同じ大きさの単位超立方体で与えられているものと仮定して，荷重の初期値が通常のように平均ゼロの乱数で与えられるとき，両極型ネットは片極型ネットに比べて収束上優位となることを述べる．この場合，中間層と出力層の双方で，分離条件は (5.8)(5.9) 式となり，図 5.2 の示す分布は，両極型ネットのとき図 5.3(a)，片極型ネットのとき図 5.3(b) と集約されることに注意されたい．

5.5.1 初期法ベクトル分布

荷重の初期値が平均ゼロの一様乱数の場合，ユニットへの入力本数 m の増大とともに，分離超平面は原点に集中することが買らによって指摘されている[40]．正規乱数の場合，このことは解析的にも示される[49]．したがって，Λ を初期法ベクトルの分布とすると，Λ の分布領域も m の増大とともに原点に漸近することになる．この場合，荷重が片極型ユニットと両極型ユニットで同じなら，その法ベクトルは R^m で同一点となり，Λ は極性によらず同じ分布となる．

図 5.4 は，初期値を一様乱数で与えたときの $m(=2,20)$ 次元の法ベクトル 30,000 個を 2 次元部分空間に投影した分布例である．図より，一様乱数の場合，初期法ベクトルは，対角線上に密に現れて，ほぼ方形状に分布することが読みとれる．また，このときの初期法ベクトルの長さについて調べた結果，$m=2$ のときの平均が $\|\bar{\boldsymbol{h}}\|=0.869$，標準偏差が $\sigma=1.212$ で，$m=20$ のとき $\|\bar{\boldsymbol{h}}\|=0.195$，$\sigma=0.116$ であった．そして，初期法ベクトルは各象限に対して同じ割合で分布することや，その分布は乱数幅の大きさによらず同じになること，また正規乱数の場合，円形状に分布することが確かめられた．



(a) $m=2$ (b) $m=20$

図 5.4: 初期荷重を一様乱数で与えた場合の初期法ベクトル分布

5.5.2 両極型ネットの収束上の優位性

図 5.3 と図 5.4 の比較から，m が小さなときは Λ と（広義）の分離条件の共通領域は広いが，m が大きくなるにつれて，初期法ベクトルが原点に漸近するため，その共通領域は少なくなるこ

とがわかる．また，活性領域は，荷重が大きいほど，また m が大きいほど，狭くなることが知られている[40]．そのため，初期時の活性領域は，m が大きくなるにつれて，個々には原点付近を通る細い帯状の領域となり，全体としては原点から放射状に広がる帯状領域を形成することになる．したがって，この活性領域と分離条件との共通領域の大きさは，ネットサイズが大きくなるにつれて，両極型の方が片極型に比べて広くなる．また，入力集合や入力定義域を通る分離超平面や活性領域の方向についても，両極型では任意の方向がとれるが，片極型では主に第1象限の方向に限られる．これらのことは，片極型の場合，入力集合や入力定義域に対して，初期時，分割が有意となる活性領域の割合が両極型に比べて低くなることを示唆している．

さらに，分離条件内の法ベクトルの長さについてみてみると，Ω_U では，入力定義域サイズ D ($=\sqrt{m}$: 超立方体の対角長）以下となるのに対して，Ω_B では，$D/2$ 以下となっている．このことは，片極型の場合，両極型に比べて，初期法ベクトルが解の構成要素となる法ベクトルに推移するまでの距離が長くなって，収束の点で不利になることを示唆している．また，これからの類推で，片極型の場合，解を構成する法ベクトルの内，大多数はその長さが $D/2$ 未満でも，いくつかは $D/2$ 以上となる解を得るのは，m の増大とともに難しくなると考えられる．事実，片極型ネットにより実現される解の種類は制約されて偏る[49]．

以上のことから，通常のように荷重の初期値を平均がゼロの乱数で設定する場合，BP 学習の収束は両極型ネットの方が片極型ネットに比べて優位になるといえる．

5.6　シミュレーションと考察

ここでは，初期値を通常のように平均ゼロの一様乱数で設定したときの BP 学習について，分類問題に対する両極型ネットと片極型ネットの収束能力をシミュレーションにより比較検討する．また，このときの両極型ネットと分割が幾何学的に等価となるように片極型ネットの初期値を $\boldsymbol{w}_U = \boldsymbol{w}_B$, $b_U = b_B - \mathbf{1}^{\mathrm{T}}\boldsymbol{w}_B/2$ と設定したとき[10]の収束能力についても検討する．さらに，初期値を分離条件内に配置することの収束上の効果について考察する．

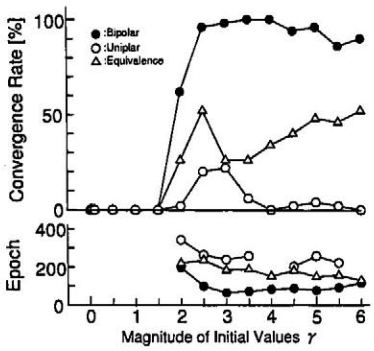


図 5.5: 8 次のパリティ問題による学習結果

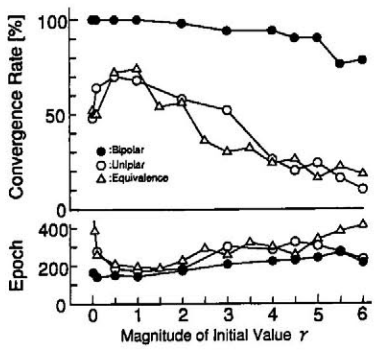


図 5.6: ソナー問題による収束挙動の違い

5.6.1 パリティ問題とソナー問題の学習結果

まず，提示入力パターンが離散的なパリティ問題（8 次）について，両極型（片極型）のとき，提示入力パターンを $\pm 0.5(0$ か $1)$ と正規化し，教師を ± 0.4 (0.1 か 0.9) と与えて，入力，中間，出力層のユニット数がそれぞれ 8, 30, 1 個のネット $(8-30-1)$ で BP 学習させた．その際，最大繰返し回数は 100,000 回，学習係数は $\eta = 0.2,\ 0.6,\ 1.0$，慣性係数は $\alpha = 0.9$ とし，初期荷重は平均ゼロの一様乱数 $[-W, W]$ $(W = 0.01,\ 0.1,\ 0.5,\ 1.0,\ 1.5, \cdots,\ 6.0)$ とした．そして，収束判定基準はすべて訓練パターンに対して誤差の絶対値が 0.1 未満になることとした．図 5.5 に $\eta = 0.6$ のときの収束率（Convergence Rate：50 回の試行に対する収束数の割合）と平均エポック数 (Epoch) を示す．図中のマーク●（○）は，両極型（片極型）ネットによる結果，また△は片極型ネットの初期値を幾何学的な分割が両極型ネットと等価となるように与えたときの結果である．図より，両極型による収束率や平均エポック数は，片極型に比べて，広い範囲の W で良好となることがわかる．また，$\eta = 0.2,\ 1.0$ の場合でも同様の結果が得られた．

つぎに，提示入力パターンが連続的なソナー問題[†]について，$60-12-1$ のネットで提示入力パターンを同様に正規化し，収束判定基準を平均 2 乗誤差が 0.01 未満になることとして学習させた．この場合，学習係数を $\eta \geq 0.6$ と大きく設定すると極性によらず全く収束しなかったため，ソナー問題では，$\eta = 0.1,\ 0.5$ について比較する．$\eta = 0.5$ のときの結果を図 5.6 に示す．図より，このときも両極型による収束は片極型に比べて優れていることがわかり，同様の結果が，$\eta = 0.1$ とした場合でも得られた．

[†]http://www.cs.cmu.edu/afs/cs/project/connect/ bench/ のベンチマークデータを使用

上述の図5.5,5.6から，片極型ネットの場合，両極型ネットと分割が等価となるように設定しても，両極型ネットによる収束結果には及ばないことが読み取れる．その他の学習係数についても同様の結果が得られた．以下，このことについて考察する．初期分割が幾何学的に等価な場合，片極型ネットの入力パターン集合は両極型ネットのそれを $\mathbf{1/2}$ だけ平行移動したものとなっており，また出力層への入力集合についても少なくとも初期時ものは $\mathbf{1/2}$ だけ平行移動したものとなっている．したがって，これらの集合は，極性によらず，いずれも合同な外接立方体や直方体で近似される．しかし，図5.2と図5.3から，分離超平面が立方体や直方体を通るときの法ベクトルは，平均的に，片極型のときの方が両極型に比べて長くなる．とくに，出力層に対する入力集合について見てみると，その外接直方体を通る分離超平面に随伴する法ベクトルは，両極型の場合，図5.2の (a), (b), (c) のいずれの分布も取り得るのに対して，片極型の場合，シグモイド関数の定義より外接直方体は原点を含まないため，図5.2(c) の環状分布 $\Omega^{\sharp}_{\mathrm{off}}$ しか取れない．$\Omega^{\sharp}_{\mathrm{off}}$ は，原点から入力集合までの距離が長くなるほど，空洞部分は広くなって，環自体の厚みも全体的に薄くなる．このことは，片極型の場合，法ベクトルの推移に要する距離が長くなるうえに，学習係数が大きくなるほど，分離条件としての環内に法ベクトルが留まる可能性が低くなることを示唆している．このため，初期分割が幾何学的に等価な場合でも，片極型による収束は両極型に比べて劣ると考えられる．

5.6.2　分離条件を考慮した初期値設定の効果

ここでは，初期法ベクトルを分離条件内に配置することの収束上の効果を調べるため，XOR問題を $2-2-1$ のネットで，最大繰り返し回数を 10,000 回，$\eta = 0.1,\ 0.2,\ 0.6,\ 1.0$ ，$W = 0.01,\ 0.1,\ 0.5,\ 1,\ 2,\ \cdots,\ 10$ として学習させた．そして試行数を1,000回とした以外は前述のパリティ問題と同条件とした．

図5.7に $\eta = 0.1$ のときの収束率と平均エポック数を示す．図中のマークは，両極型（片極型）ネットで分離条件を考慮したときの結果◆（◇）と考慮しなかったときの結果●（○）を表す．分離条件を考慮したときの収束率は，◆と◇の比較から，収束率と平均エポック数ともに，$W \leq 0.5$ では片極型の方が，また $W \geq 1$ では両極型の方が良好な結果となっている．そして，分離条件

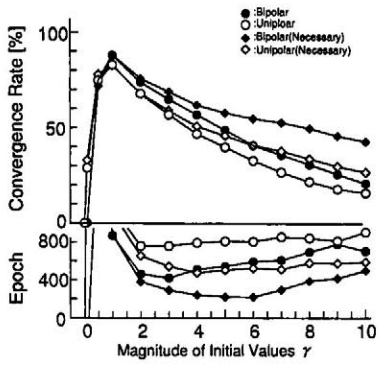


図 5.7: 分離条件を考慮した初期値設定の効果

を考慮することの効果は W が大きいとき顕著である．しかし，W が小さいときはほとんど見られない．これは，W が小さい場合，活性領域が広くなって，分離超平面の境界としての役割が薄れるため，分離条件内に初期法ベクトルを配置しても，その効果は少なくなるためであると考えられる．その他の学習係数を用いた場合でも，分離条件を考慮したときの収束率や平均エポック数は両極型ネットによる結果◆の方が優れていた．

5.7 結言

本章では，第1中間ユニットの荷重のなす分離超平面が入力集合の近似直方体の内部を通るための分離条件を法ベクトルの集合で (5.5) 式のように定式化した．また，第2中間以降の層のユニットに対する広義の分離条件を両極型ネットのときと片極型ネットのときに分けてそれぞれ (5.8) 式と (5.9) 式で与えた．そして，分離条件の高次元ベクトル空間における幾何学的形状に関する知見を得るため，2次元のときの分離条件を図示し，入力パターン集合が単なる平行移動の関係にある場合や，ネット極性が異なる場合，その存在領域が異なることを具体的に述べた．また，荷重の初期値を通常のように平均がゼロの乱数で与えたときの初期法ベクトルの分布や初期活性領域の形状を明らかにした．そして，この結果と分離条件を対比することにより，通常の初期値設定による場合，片極型ネットに比べて，両極型ネットによる BP 学習の収束は優位となることを述べた．さらに，シミュレーションにより，分割が等価となるように初期値設定した場合でも，片極型ネットに比べて，両極型ネットによる収束は幅広い学習係数に対して良好となることを示し，

その原因ついて言及した．最後に，分離条件を考慮した初期値設定の効果について考察した．

両極型ネットによるBP学習の収束は，分類問題だけでなく近似や時系列予測問題でも優れていることが実験的に確認されている．このことを裏付けるには，分離超平面だけでなく活性領域を含む詳細な議論が必要となる．

第6章　三層ネットの初期値設定法

6.1　緒言

バックプロパゲーションネットワーク (BP ネット) は，パターン認識や制御等の分野で広く応用されており，現実には，三層構成のネットでも十分有用であることが多い[12),22)]．しかし，BP ネットでは，収束解がローカルミニマに陥いることや収束速度が遅いといった収束上の問題点が指摘されており，その運用は試行錯誤的である[3)]．これは，BP 学習が降下原理に基づくため，収束の可否が荷重の初期値に大きく依存することによる．この初期値について，賈らは，片極型ネットを対象に，初期時の分離超平面を入力定義域の中心に配置させ，活性領域幅と入力定義域を整合させるように荷重を設定する指針を提唱し，良好な収束結果を得ている[40)]．この指針の背景には，活性領域と入力集合との共通部分が広ければ，シグモイド関数の微分が微小となる可能性はなくなって，荷重の更新が円滑に行われるということがある．

一方，シグモイド関数が 0 から 1 の値をとる片極型ユニットからなる片極型ネットと −0.5 から 0.5 の値をとる両極型ユニットからなる両極型ネットは，構造が同じなら同一の解濃度を持つ[50)]．これにも関わらず，荷重の初期値を通常のように平均がゼロの一様乱数で設定した場合（以降，通常型乱数初期値），BP 学習の収束は，両極型ネットの方が片極型に比べて優れていることが多数報告されている[16),33),49)]．このことは，上述の観点からも説明できるが，より厳密には，初期法ベクトルの分布と入力定義域を分離する法ベクトルの集合を比較することにより，両者の共通領域は片極型ネットに比べて両極型ネットの方が広くなることから，説明できる[43)]．しかし，一般に各ユニットの入力集合は入力定義域よりも小さくなる傾向があるため，入力定義域の中を通るように分離超平面を配置したとしても，入力集合を通るとは限らない．

本章では，パターン分類問題を三層両極型 BP ネットで学習させる際の初期値設定法について

考察を行う．すなわち，中間層については，そのユニットのなす分離超平面が提示入力そのものの集合の中心を通るように荷重を設定し，出力層については，先験情報が偏って与えられないように，すべての荷重をゼロと設定する．パリティ問題やMONK's 問題，ソナー問題，アヤメの分類問題に対するシミュレーションの結果，提案法により良好な収束性が得られて，提案法が有効であることを確認した．

6.2　入力集合と入力定義域

ここでは，入力ユニット数が m 個，中間ユニット数が n 個，出力ユニット数が l 個の三層からなるネット $(m-n-l)$ で，提示入力と教師の対がそれぞれ $\boldsymbol{x}_s = [x_1^s, x_2^s, \cdots, x_m^s]^T$ と $\boldsymbol{d}_s = [d_1^s, d_2^s, \cdots, d_l^s]^T$ $(s = 1, 2, \cdots, S)$ で与えられる分類問題を学習させる場合を考える．ここに，$\boldsymbol{d}_s$ の要素は ± 0.5 のように二値をとるものとする．このとき，提示入力 $\boldsymbol{x}_s$ に対して，中間層の第 $j(=1, 2, \cdots, n)$ 番目のユニットでの総入力 u_j^s と出力 y_j^s は，(2.1)(2.2) 式より

$$u_j^s = \sum_{i=1}^{m} x_i^s w_{ji} + b_j \tag{6.1}$$

$$y_j^s = f(u_j^s) = \frac{1}{2}\tanh(\frac{u_j^s}{2}) \tag{6.2}$$

と書き改められる．また，出力ユニットについても，その結合荷重とバイアス荷重をそれぞれ v_{kj} と c_k として，総入力 q_k^s と（ネット）出力 z_k^s が同様に計算されるものとする．

いま，第 j 番目の中間ユニットの形成する分離超平面 $\{\boldsymbol{x} \in R^m | \boldsymbol{w}_j^T \boldsymbol{x} + b_j = 0\}$ を H_j と表記することにする．ここに，R^m は m 次元ユークリッド空間である．このとき，R^m は H_j を境に高出力域と低出力域に分かれ，さらに，その各々は活性領域 $\mathcal{A}_j = \{\boldsymbol{x} \mid -0.5 + \rho < f(\boldsymbol{w}_j^T \boldsymbol{x} + b_j) < 0.5 - \rho\}$ と飽和領域 $\mathcal{S}_j = \{\boldsymbol{x} \mid f(\boldsymbol{w}_j^T \boldsymbol{x} + b_j) < -0.5 + \rho,\ f(\boldsymbol{w}_j^T \boldsymbol{x} + b_j) > 0.5 - \rho\}$ に分かれる．そして，活性領域の幅 g_j や法ベクトル $\boldsymbol{h}_j$ は，(2.17)(2.18) 式より

$$\boldsymbol{h}_j = -\frac{b_j}{\|\boldsymbol{w}_j\|^2}\boldsymbol{w}_j \tag{6.3}$$

$$g_j = \frac{2\ln\{(1-\rho)/\rho\}}{\|\boldsymbol{w}_j\|} \tag{6.4}$$

となる．同様に，出力ユニットについても，分離超平面 H_k，法ベクトル $\boldsymbol{h}_k$，活性領域 A_k やそ

の幅 g_k が定義できる．以下では，入力パターン集合を $\mathcal{X} = \{\boldsymbol{x}_s\}$，$\mathcal{X}$ に外接する超直方体を入力パターン定義域 $\boldsymbol{X}$，中間層からの出力集合を $\mathcal{Y} = \{\boldsymbol{y}_s = [y_{s1}, y_{s2}, \cdots, y_{sn}]^T\}$ と定義して，中間層と出力層に対する入力集合の性格の違いを調べる．まず，提示入力のとる値は，学習課題に依存するものの，その範囲については特に制限がなく，提示入力集合 $\mathcal{X}$ が学習期間中変わることはない．一方，出力ユニットに対する入力 $\boldsymbol{y}_s$ については，シグモイド関数の規定する n 次元超立方体 (入力空間) $\boldsymbol{Y} = [-0.5, 0.5]^n$ 内の値をとって，出力層での入力集合は $\mathcal{Y} \subset \boldsymbol{Y}$ のように $\mathcal{Y}$ の部分集合となる．そして，出力層に対する入力集合 $\mathcal{Y}$ は，中間層での荷重更新によって学習期間中たえず変化する．

したがって，中間ユニットについては，提示入力集合を最終的に定まる入力集合として取り扱えるが，出力ユニットについては，最終的に確定する入力集合を予め知ることはできない，という違いがある．

6.3 初期値設定法

前節で述べたように，中間ユニットに対する入力集合と出力ユニットに対する入力集合はその性質が異なる．そこで，本節では，中間ユニットに対しては，分離超平面が入力集合の中心を通るように設定し，出力ユニットに対しては，先験情報を与えないように全ての荷重をゼロで設定する初期値設定法について議論する．

6.3.1 中間層に対する初期値設定

パターン分類問題に対してネットが解を持つためには，少なくとも幾つかのユニットのなす分離超平面が入力集合の内部を通る必要がある．この観点から，5.3.1 節では，入力集合に外接する超直方体を求めて，その内部を分離超平面が通るための分離条件を法ベクトルの集合 Ω として定式化した[43]．そして，通常のように荷重の初期値を平均がゼロの一様乱数で与えた場合，初期法ベクトルの集合 Ψ_0 と Ω の共通領域は，片極型ネットに比べて両極型ネットの方が広くなること明らかにして，両極型ネットによる収束が優位となることを裏付けた[43]．しかし，上記のように

初期値を一様乱数で与えた場合，図 6.1 に示すように，必ずしも全ての分離超平面 H_j が提示入力集合 $\mathcal{X}$ の内部を通るとは限らない．

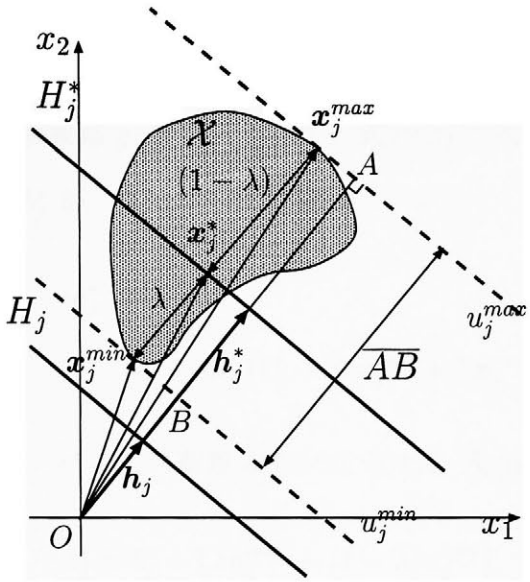


図 6.1: 入力集合の内部を通る分離超平面

そこで，以下では，分離超平面 H_j を平行移動して，新たに得られる分離超平面 H_j^* が提示入力集合 $\mathcal{X}$ を通るようにする．

提示入力 $\boldsymbol{x}_s$ に対する中間ユニットへの総入力 u_j^s の集合 $\{u_j^s | s = 1, 2, \cdots, S\}$ には最大値 u_j^{max} と最小値 u_j^{min} が存在する．

この u_j^{max} と u_j^{min} を与える入力 $\boldsymbol{x}_s$ をそれぞれ

$$u_j^{max} = \boldsymbol{w}_j^T \boldsymbol{x}^{max} + b_j \tag{6.5}$$

$$u_j^{min} = \boldsymbol{w}_j^T \boldsymbol{x}^{min} + b_j \tag{6.6}$$

のように $\boldsymbol{x}_j^{max}$ と $\boldsymbol{x}_j^{min}$ で表す．そして，$\boldsymbol{x}_j^{min}$ と $\boldsymbol{x}_j^{max}$ を λ 対 $(1-\lambda)$ に内分する点

$$\boldsymbol{x}^* = (1-\lambda)\boldsymbol{x}_j^{min} + \lambda \boldsymbol{x}_j^{max} \tag{6.7}$$

を通って，H_j に平行な分離超平面を新たに H_j^* と定義する．以降，λ を内分比とよぶことにする．

この場合，$0 \leq \lambda \leq 1$ ならば，H_j^* は必ず提示入力集合 $\mathcal{X}$ を通ることになる．また，H_j^* に随伴する法ベクトル $\boldsymbol{h}_j^*$ は

$$\boldsymbol{h}_j^* = \frac{\boldsymbol{h}_j^T\{(1-\lambda)\boldsymbol{x}_j^{min} + \lambda \boldsymbol{x}_j^{max}\}}{\|\boldsymbol{h}_j\|^2}\boldsymbol{h}_j \tag{6.8}$$

と導かれる．のとき，活性領域幅 g_j を同一として，式(6.8)に式(6.3)を代入すれば，分離超平面 H_j^* を構成する荷重 $\boldsymbol{w}_j^*$ と b_j^* が

$$\boldsymbol{w}_j^* = \boldsymbol{w}_j \tag{6.9}$$

$$b_j^* = -\boldsymbol{w}_j^T\{(1-\lambda)\boldsymbol{x}_j^{min} + \lambda \boldsymbol{x}_j^{max}\} \tag{6.10}$$

のように得られる．さらに，バイアス荷重 b_j^* については，式(6.5)(6.6)より，

$$b_j^* = b_j - \{\lambda u_j^{max} + (1-\lambda)u_j^{min}\} \tag{6.11}$$

のように総入力の最大値 u_j^{max} と最小値 u_j^{min} から簡単に算出できる．

入力集合 $\mathcal{X}$ と活性領域 $\mathcal{A}_j$ との共通領域が広ければ収束能力は高くなることが確かめられている[43]．そこで，収束能力の観点から内分比についてみてみると，$\lambda = 0.5$ の場合，入力集合 $\mathcal{X}$ の中心を通ることになって，入力集合 $\mathcal{X}$ と活性領域との共通領域は，最も広くなる．一方，$\lambda = 0.0, 1.0$ の場合，分離超平面は入力集合 $\mathcal{X}$ の両端を通ることになって，$\lambda = 0.5$ とした場合に比べて共通領域は狭くなる．したがって，内分比は $\lambda = 0.5$ とする．このことの妥当性は後述のシミュレーションで示す．

以上のことから，中間ユニットの結合荷重の初期値 $\boldsymbol{w}_j^*$ については，式(6.9)のように通常型乱数初期値として与えられる $\boldsymbol{w}_j$ をそのまま用い，バイアス荷重の初期値 b_j^* については，$\lambda = 0.5$ として，式(6.11)のように設定し直す．

6.3.2 出力層に対する初期値設定

出力ユニットに対する入力定義域 $\boldsymbol{Y}$ は，出力ユニットの荷重がゼロでない場合，分離超平面を境に高出力域と低出力域に分かれ，どちらが高出力域になるか低出力域になるかはバイアス荷

重の符号で一意に決まる．すなわち，$\boldsymbol{v}_k \neq \boldsymbol{0}$, $c_k \neq 0$ の場合，H_k からみて原点側の点 $\boldsymbol{y}$ では，

$$\boldsymbol{h}_k^T(\boldsymbol{y}-\boldsymbol{h}_k) = -\frac{c_k q_k}{\|\boldsymbol{v}_k\|^2} < 0 \tag{6.12}$$

となって，c_k が正（負）ならば高（低）出力域となる[51]．

一方，出力ユニットの初期荷重を $\boldsymbol{v}_k = \boldsymbol{0}$, $c_k = 0$ で与えた場合，$\boldsymbol{Y} \subset H$ となって，分離超平面は入力集合全体を覆うことになる．これは，全ての提示入力に対して，ネット出力は全てゼロとなって，高低差が無くなることを意味している．

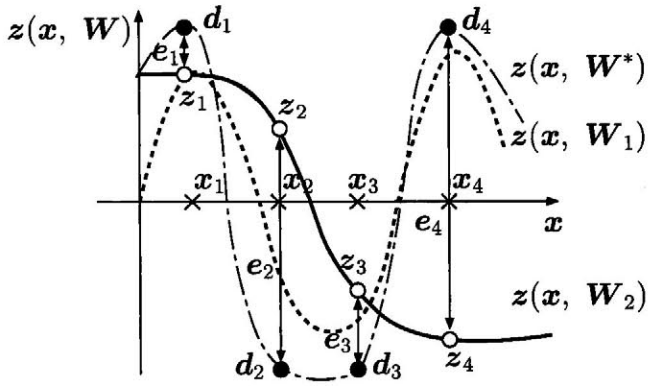


図 6.2: 多層ネットの出力曲面の概念図

図 6.2 は，訓練データの入力パターンと教師の対が $(\boldsymbol{x}_s, \boldsymbol{d}_s)$ $(s = 1, 2, 3, 4)$ と 4 個与えられる場合について，ネットの入出力関係を概念的に表したものである．図中，横軸は $\boldsymbol{x} \in R^m$ なる任意の提示入力で，縦軸はネット内のすべての荷重を $\boldsymbol{W}$ と表記したときの $\{\boldsymbol{x}\}$ に対するネット出力 $\boldsymbol{z}(\boldsymbol{x}, \boldsymbol{W})$ である．$\boldsymbol{z}(\boldsymbol{x}, \boldsymbol{W})$ は超曲面をなしており，$\boldsymbol{z}_s = \boldsymbol{z}(\boldsymbol{x}_s, \boldsymbol{W})$ となって，教師との誤差は $\boldsymbol{e}_s = \boldsymbol{d}_s - \boldsymbol{z}_s$ となることに注意されたい．また，$\boldsymbol{W}^*$ はある一つの収束解，$\boldsymbol{W}_1$ と $\boldsymbol{W}_2$ はそれぞれ通常の乱数による初期荷重を表している．初期値が $\boldsymbol{W}_1$ で与えられて，学習開始時のネット出力が $\boldsymbol{z}(\boldsymbol{x}_s, \boldsymbol{W}_1)$ のように教師 $\boldsymbol{d}_s$ と適合している場合，以降での学習は円滑に進むと考えられる．この場合，初期値 $\boldsymbol{W}_1$ は訓練データを適切に反映した先験情報をネットに与えていると考えられる．しかし，現在のところ，このように適切な先験情報を抽出する方法は見当たらない．また，初期値を通常のように平均ゼロの小さな乱数で与えた場合，ユニットへの入力数 (m や n) が

増加するにつれて，各層に対する入力定義域のサイズは $\sqrt{m}$ または $\sqrt{n}$ に比例して増大するのに対して，活性領域幅は逆に式(6.4)のように狭くなるため，飽和領域に含まれる入力が増えてくる．

図6.2のネット出力 $\boldsymbol{z}(\boldsymbol{x}, \boldsymbol{W}_2)$ は，飽和領域に含まれる入力が増えたときの様子を表わしたものである．すなわち，この例で $\boldsymbol{x}_2$ や $\boldsymbol{x}_3$ のような提示入力については，$\boldsymbol{y}_2$ や $\boldsymbol{y}_3$ が，出力ユニットからみて，シグモイド関数の微分が比較的大きな活性領域にあることになって，荷重は更新される．一方，$\boldsymbol{x}_1$ や $\boldsymbol{x}_4$ のような提示入力については，$\boldsymbol{y}_1$ や $\boldsymbol{y}_4$ は，微分がほぼゼロの飽和領域あることになって，荷重は更新されない．ただし，学習を収束させるためには，$\boldsymbol{x}_1$ のときのようにネット出力と教師との誤差が非常に小さい提示入力の他に，$\boldsymbol{x}_4$ のときのようにネット出力が教師と反対側に飽和して誤差が非常に大きな提示入力についても，最終的にネット出力と教師との誤差を小さくする必要がある．そのため，学習開始時の誤差が大きい訓練データについては，それ自身を提示しても荷重が更新されないため，開始時の大きな誤差はその他の提示入力による荷重更新よって小さくするしかない．しかし，その他の提示入力よる荷重更新によって，誤差が小さくなる保証は必ずしもない．したがって，通常のように小さな乱数を初期値として与える場合でも，ネットサイズが大きくなるにつれて，学習の収束は難しくなると考えられる．

一方，出力ユニットの初期荷重をすべてゼロで設定した場合，前述のように $\boldsymbol{Y} \subset H$ となって，$\boldsymbol{y}_s$ はすべて活性領域 $\boldsymbol{A}_k$ にあることになる．したがって，出力ユニットの荷重は，すべての提示入力に対して，更新されることになる．また，このときのネット出力はすべてゼロとなって，教師との誤差の絶対値はどの提示入力に対しても等しく0.5となって偏ることはない．出力ユニットの荷重は，具体的に，

$$\Delta v_{kj} = \eta f'(q_k^s) e_k^s y_j^s \tag{6.13}$$

のように更新される．ここに，η は学習係数である．この場合，学習開始時の荷重更新は，教師とネット出力との誤差が $e_k^s = d_k^s - z_k^s = d_k^s$ となって教師の影響を強く受ける．さらに，中間ユニットの荷重は，

$$\Delta w_{ji} = \eta f'(u_j^s) \sum_{k=1}^{l} v_{kj} f'(q_k^s) e_k^s y_j^s \tag{6.14}$$

のように更新される．上式の v_{kj} は，学習開始時でも，出力ユニットの荷重更新後，Δv_{kj} となって必ずしもゼロとはならないため w_{ji} も学習開始時から更新されることに注意されたい．

したがって，出力ユニットの荷重の初期値はつぎのようにすべてゼロと設定する．

$$\boldsymbol{v}_k = \mathbf{0} \tag{6.15}$$

$$c_k = 0 \tag{6.16}$$

以上に述べた初期値設定法について，その具体的な手順を示す．

Step 1 中間ユニットの荷重 $\boldsymbol{w}_j, b_j$ を一様乱数で初期化する．

Step 2 提示入力 $\boldsymbol{x}_s$ を入力し，中間ユニットへの総入力の最大値 u_j^{max} と最小値 u_j^{min} を求める．

Step 3 u_j^{max} と u_j^{min} を基に，内分比 λ として，式(6.11)により，中間ユニットの結合荷重はそのままで，バイアス荷重 b_j^* を再設定する．

Step 4 出力ユニットの荷重を $\boldsymbol{v}_k = \mathbf{0}$, $c_k = 0$ と設定する．

このような手順で初期値設定することにより，中間ユニットでは分離超平面が入力集合を通るように初期荷重が再設定され，出力ユニットでは初期荷重がゼロと設定される．ただし，内分比 λ については，どの程度の値が適切かを以降のシミュレーションにおいて調べる．

6.4　シミュレーション

はじめに，内分比を $\lambda = 0.5$ と定めることの妥当性を検証するため，6次のパリティ問題(6–12–1)について，内分比を範囲が [0, 1] の一様乱数や，$\lambda = 0.0,\ 0.5,\ 1.0$ で与えた場合の収束能力をシミュレーションにより比較した．BP学習は，学習係数を $\eta = 0.01,\ 0.05,\ 0.1$，慣性係数を $\alpha = 0.9$，最大繰返し回数を 100,000 回，収束条件を全ての訓練データに対する出力誤差の絶対値が 0.1 未満となることとした．収束性は，50 回の試行をもとに収束率（50 回のうちの収束した試行数の割合），平均学習回数（収束に至るまでの平均の Epoch 数）で評価した．このときの学習結果を図6.3に示す．図中，横軸の γ は通常型乱数初期値の幅，すなわち一様乱数の幅 $[-\gamma, \gamma]$ $(\gamma = 0.01, 0.1, 0.5, 1.0, 1.5, \cdots, 6.0)$ を表し，縦軸はそれぞれ収束率と平均学習回数を表している．

図6.3は6次のパリティ問題を $\eta = 0.1$ として学習させた結果である．そして，∇ $(\lambda = 0)$ と $\triangle$ $(\lambda = 1)$ は分離超平面が入力集合の両端に接する場合，○は分離超平面が入力集合の中心を通る

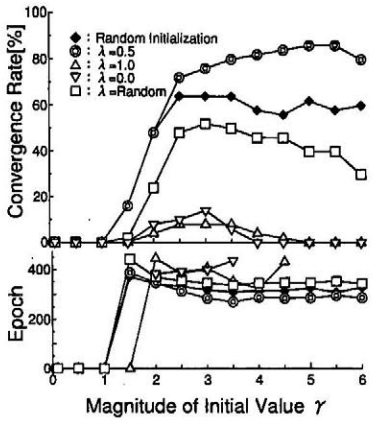


図 6.3: 内分比による収束性の違い

場合 ($\lambda = 0.5$)，□は，内分比 λ を一様乱数で与えた場合，◇は乱数初期値を用いた場合の学習結果である．図 6.3 からわかるように，内分比を $\lambda = 0.5$ としたときの学習結果は他に比べて優れており，乱数初期値のときに比べて，$\gamma = 5.0$ による収束率で 32%，学習回数で 93 Epoch の改善がみられる．また，乱数初期値による収束率は，$\gamma = 3$ でのピークを境に，γ の増加とともに減少している．これに対して，内分比を $\lambda = 0.5$ とした場合，$2 < \gamma < 5.5$ と広い範囲で 70% 以上の収束率を維持している．一方，内分比 λ を 0 または 1 とした場合，ほとんど収束しないことが読みとれる．そして，λ を乱数で与えたときの収束率は，30% から 60% 程度で乱数初期値のときより悪い．さらに，学習係数を $\eta = 0.5, 1.0$ とした場合についてもシミュレーションを行ったところ，内分比を $\lambda = 0.5$ としたときの収束は乱数初期値のときに比べて良好で，収束性が改善されることを確認した．また，内分比を $\lambda = 0, 1$ としたときの収束率は学習係数が大きくなるにつれて向上する傾向がみられた．

以上のことから，収束性については，提案する初期値設定法において，$\lambda = 0.5$ として，分離超平面が入力集合の中心を通るように初期値設定したときに，最も良好な収束結果が得られて，内分比を $\lambda = 0.5$ とすることが妥当であることが分かる．また，この妥当性はソナー問題でも確かめられた．

次に，提案法の有効性を確認するため，アヤメの分類問題 (4–3–3，連続値入力，30 組の訓練データ)，ソナー問題 (60–4–1，連続値入力，104 組の訓練データ)，MONK'S 問題 (6–4–1，離散値

入力，124組の訓練データ) を対象に，中間層と出力層の初期荷重を表6.1のように与えて，平均自乗誤差が0.01未満となるまでの収束性を比較した．表中，○は中間層の初期値を(6.9)(6.11)式で

表 6.1: 中間層と出力層の初期荷重設定

	$\boldsymbol{w}_j,\ b_j$	$\boldsymbol{v}_k,\ c_k$
◆	Random	Random
◇	Random	$\boldsymbol{v}_k = 0,\ c_k = 0$
●	(6.9)(6.11) 式	(6.9)(6.11) 式
◎	(6.9)(6.11) 式	Random
○	(6.9)(6.11) 式	$\boldsymbol{v}_k = 0,\ c_k = 0$

◆:通常乱数初期値, ○:提案法

与えて出力層の荷重をすべてゼロで設定する提案法を，◆は通常型乱数初期値設定 (“Random”) を表している．また，中間層と出力層に対する初期値設定の効果を調べるために，通常型乱数初期値で出力層の初期値をゼロで与えた場合（◇），中間層の初期値は(6.9)(6.11)式で与え，出力層の初期荷重を通常型乱数初期値とした場合（●），全ての層に対して初期値を(6.9)(6.11)式で与えた場合（◎）についても実験を行った．以上の結果を図6.4から図6.6に示す．

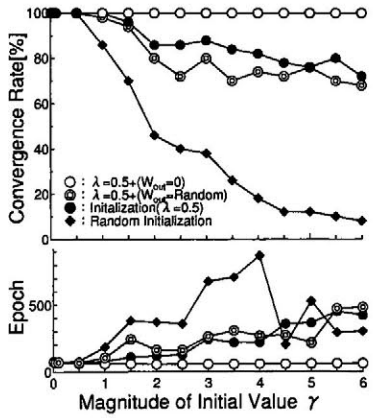


図 6.4: アヤメの分類問題の学習結果

図6.4はアヤメの分類問題の学習結果を表している．通常型乱数初期値を用いた場合の結果◇

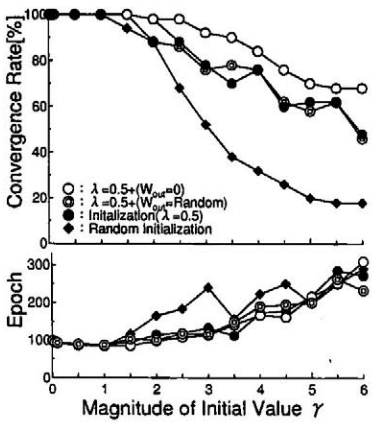


図 6.5: MONK'S 問題の学習結果

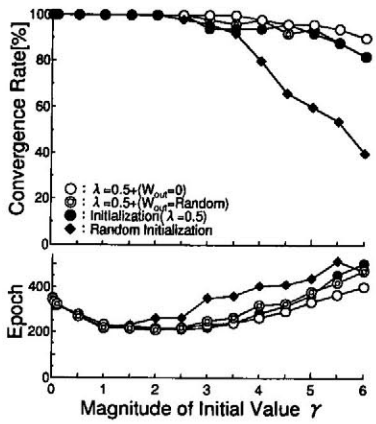


図 6.6: ソナー問題の学習結果

は乱数幅が大きくなるにつれてほとんど収束しなくなっているのに比べて，中間ユニットの分離超平面が提示入力集合の中心を通るように初期値設定（●，◎，○）することで，乱数初期値に比べて収束性の改善が図られることが分かる．そして，提案法（○）では，初期荷重幅に依らず100%の収束率を示しており，その他の方法がγが大きくなるにつれて収束率が低下しているのに比べても良好な結果が得られていることが読み取れる．また，$\eta = 0.01,\ 0.1$ とした場合でも，提案法による収束性は他に比べて良好であった．

次に，MONK's 問題 の学習結果を図 6.5 に示す．MONK's 問題の学習においても乱数初期値による結果◆に比べて提案法を適用した場合の結果○では，収束率は$\gamma = 3.5$ のとき 40%改善され，平均学習数も若干の改善されていることが分かる．また，アヤメの分類問題と同様，○と◎，◆の収束率は，ほぼ同程度であることが読み取れる．

図 6.6 はソナー問題の学習結果を表している．ソナー問題においても通常型乱数初期値による結果◆に比べて，提案法による結果○では収束率は$\gamma \geq 5.0$ のとき 50%の改善が見られ，平均学習回数も他の手法に比べて少ないことが読み取れる．また，中間層の初期値を乱数型初期値で与えた場合，$\gamma \geq 3.5$ の範囲で収束率が急速に落ち込んでいるが，初期値を (6.9)(6.11) 式で与えた場合，初期荷重幅に拠らず良好な収束率を示していることが分かる．

以上の結果から，中間層に対しては，分離超平面が入力集合の中心を通るように (6.9)(6.11) 式によって設定し，出力層の初期荷重については全てゼロで設定することで，通常型乱数初期値に比べて収束性の改善が図られることが分かった．また，(6.9)(6.11) 式を出力層に対して適用する

ことは，悪影響は及ぼさないものの収束性にはあまり影響しないといえる．さらに，図6.4のアヤメの分類問題の学習結果から，収束率は，乱数幅によらず，ほぼ100%で，学習回数も同程度となっており，少なくともアヤメの分類問題においては，初期値依存性が緩和されている．

6.5　結言

本章では，パターン分類問題を両極型BPネットで学習させる際の初期値設定法を提案した．すなわち，中間層については，そのユニットのなす分離超平面が提示入力集合の中心を通るように荷重を設定し，出力層については，先験情報が偏って与えられないように，すべての荷重をゼロと設定した．パリティ問題やMONK's問題，ソナー問題，アヤメの分類問題に対するシミュレーションの結果，提案法により良好な収束性が得られて，提案法が有効であることを確認した．

従来，入力パターンをクラスタリングする既存法の結果を中間ユニットの初期荷重の設定に反映できたとしても，出力ユニットの初期荷重は乱数で与えられているため，クラスタリングの効果が収束にどのように影響するか分からなかった．しかし，前節のシミュレーション結果から，BP学習は中間ユニットの初期荷重をある程度大きな値で与えた場合でも出力ユニットの初期荷重をゼロで設定すれば，良好な収束性が得られていることが分かる．このことは，三層ネットの初期値設定問題が中間層における初期値設定問題に帰着できる可能性を示している．

第7章　結論

本論文では，これまで定性的にしか議論されていなかった収束性に関する議論を分離超平面に随伴する法ベクトルの概念を導入して幾何学的観点から考察を行った．すなわち，幾何学的観点からみた学習収束を分離超平面が入力集合を有意義に分割することであると捉え，これを多層ネットが解を持つための必要条件とみなし，この必要条件と初期値を解析することで解のアトラクタの観点から静的な学習特性に関する議論を展開した．その際，入力集合は学習課題に依存しており，仮に入力集合が特定できても，その内部を有意義に分離する分離超平面を特定することは難しいため，解の必要条件を厳密に導出することは困難となるため，必要条件に代わる近似的な条件として，分離超平面が入力集合の外接直方体の内部を通る条件が求まれば，収束性に関する厳密な議論が可能となる．この近似条件を分離条件として法ベクトルの集合により定式化した．そして，2次元の場合の両極型ネットと片極型ネットの分離条件を法ベクトルの分布として図示することで，分離条件の高次元ベクトル空間における幾何学的な形状に関する知見が得られた．また，入力パターン集合が単に平行移動した場合やネット極性が違う場合でも，分離条件は大きく異なることを明らかにした．一方，多層ネットの初期時の法ベクトル分布についても調べ，ユニットへの入力本数が多くなれば法ベクトルや分離超平面はユークリッド空間の中心に漸近することを明示した．そして，分離条件と初期値を同一ベクトル空間上で比較することで，これまで定性的にしか示されていなかった両極型ネットの片極型ネットに対する収束上の優位性を裏付けた．また，両極型ネットと片極型ネットの許容解濃度に関して検討し，解濃度は極性に拠らず提示入力パターンがアフィン変換の関係にあれば同一となり，分離能力も等しくなることを明らかにした．この結果をもとに，多層ネットをハードウェア化する際の設計指針を提案した．さらに，法ベクトルの概念に基づいた三層ネットの初期値設定法について提案した．すなわち，両極型ネットを対象として，中間層に対しては，分離超平面が入力集合の中心を通るように初期荷重を設定し，

出力層に対しては，先験情報を与えないように全ての初期荷重をゼロで設定する方法を提案した．そして，シミュレーションにより提案する初期値設定法を用いることで通常の乱数初期値に比べて収束性が改善されることを示した．また，出力ユニットに対する荷重設定を必要としないため，三層ネットの初期値設定問題が中間層における初期値設定問題に帰着できる可能性を示している．

多層ネットの収束性を幾何学的に解析することで，その本質的な収束性改善へ向けての基本方針が打ち出せたと考えられる．すなわち，まず，ユニット特性については，従来の片極型ユニットに代えて両極型ユニットを用いる．そして，ネットワークの各層における初期荷重については，中間ユニットでは分離超平面が入力集合の中心を通るように設定し，出力ユニットでは全ての荷重をゼロで初期化することで収束性の向上が図られると考えられる．また，このようにして得られた多層ネットを実際にハードウェア化する際に生じるスケーリング問題に対しても，4章で述べたスケーリング手法を用いることで等価な汎化能力を持つハードウェアの構築が可能となる．なお，ここで得られた結果は，リカレントネットや相互結合型ネットに対しても有効であると考えられる．

謝辞

平成８年本学大学院情報工学研究科に入学し，本研究に着手して以来今日に至るまで，本学制御システム工学教室の熊丸耕介教授には終始懇切な御指導と御討論を頂きました．また，近畿大学九州工学部の五反田博教授には学部時代から研究に関する御助言や御討論を頂きました．そして，本学制御システム工学教室の石川真澄教授をはじめ安井湘三教授，機械システム工学教室の安部憲広教授には貴重な御意見や御鞭撻を頂きました．ここに謹んで諸先生方に対して感謝の意を表します．

さらに，本研究を遂行するにあたり，色々と貴重な御助言や御援助を賜った本学制御システム工学科の井上勝裕助教授ならびに御指導や御助言を頂いた全ての方々に深く感謝いたします．

参考文献

[1] D.E.Rumelhart andd J.L.McClelland. *Parallel Distributed Processing*, Vol. 1. MIT Press., 1969.

[2] A.V.Ooyen. Improving the convergence of the back-propagation algorithm. *Neural Networks*, Vol. 5, No. 3, pp. 465–471, 1992.

[3] J. P. Bigus. *Data Mining with Neural Networks.* The McGraw-Hill Companies, 1996.

[4] Y.Le Cun, J.S.Denker, and S.A.Solla. Optimal brain damage. *Neural Information Proceeding Systems*, Vol. 2, pp. 598–605, 1990.

[5] D.E.Rumelhar, G.E.Hinton, and R.J.Wiliams. Learning representations by back-propagating error. *Nature*, Vol. 323, No. 9, pp. 533–536, 1986.

[6] D.Nguyen and B.Widrow. Improving the learning speed of two-layer neural networks by choosing initial values of the adaptive weights. *Proc.International Joint Conference on Neural Networks*, Vol. 3, pp. 21–26, 1990.

[7] B.K. Dolenko and H.C. Card. Tolerance to analog hardware of on-chip learning in backpropagation networks. *IEEE Trans. Neural Networks*, Vol. 6, No. 5, pp. 1045–1052, 1995.

[8] S. I. Gallant. *Neural Network Learning and Expert Systems.* MIT Press., 1993.

[9] G.Cybenko. Approximation by superposition of a sigmoid function. *Signals and Systems*, Vol. 2, pp. 183–192, 1989.

[10] H. Gotanda, Y. Ueda, H. Shiratsuchi. Solution space and bp learning behaviour of multilayer networks whose units are different in polarity. *J. Robotics and Mechatronics*, Vol. 7, No. 4, pp. 336–343, 1995.

[11] Akaike H. A new look at the statistical model identification. *IEEE Trans. Automatic Control*, Vol. AC-19, No. 6, pp. 716–723, 1974.

[12] S. Haykin. *Neural Networks – A Comprehensive Foundation.* Macmillan Publishing Company, 1994.

[13] R. A. Jacobs. Increased rate of convergence through learning rate adaption. *Neural Networks*, Vol. 1, pp. 295–307, 1988.

[14] J.P.Cater. Successfully using peak learning rates of 10 (and greater) in back-propagation networks with the heuristic learning algorihtm. *Proc. IEEE International Conference on Neural Network*, pp. 645–661, 1987.

[15] J.Sietsma and R.J.F.Dow. Neural net pruning-why and how. *Proc. IEEE International Conference on Neural Networks*, Vol. 1, pp. 325–333, 1988.

[16] Niclaos B. Karayniannis. Accelerating the training of feedforward neural networks using generalized hebbian rules for initializing the internal representations. *IEEE Trans. Neural Networks*, Vol. 7, No. 2, pp. 419–426, 1996.

[17] K.Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural Networks*, Vol. 2, pp. 183–192, 1989.

[18] K.Kameyama and Y.Kougi. Neural network pruning by fusing hidden layer units. *IEICE*, Vol. E74, pp. 4198–4204, 1991.

[19] J.A. Lansner and T. Lehmann. An analog cmos chip set for neural networks with arbitrary topologies. *IEEE Trans. Neural Networks*, Vol. 4, No. 3, pp. 441–444, 1993.

[20] Y. Lee, D. Oh, and M. Kim. The effect of initial weights on premature saturation in back-propagation learning. *Proc. International Joint Conference on Neural Networks*, pp. 765–770, 1991.

[21] Ljung and Söderström. *Theory and Practice of Recursive Identification.* 1983.

[22] F. Luo and R. Unbehauen. *Applied Neural Networks for Signal Processing.* Cambridge University Press, 1997.

[23] M.A.C. Maher, S.P. Deweerth, M.A. Mahowald, C.A. Mead. Implementing neural architecures using analog vlsi circuits, 1989.

[24] M. Minsky and S. Papert. *Perceptrons.* MIT Press, 1969.

[25] M.Ishikawa. Structural learning with forgetting. *Neural Networks*, Vol. 9, No. 3, pp. 509–521, 1996.

[26] P.W Munro. Visualizations of 2-d hidden unit space. *Proc. IJCNN*, pp. III 468–473, 1993.

[27] N.B.Karayiannis and A.N.Venetsanopoulos. Fast learning algorithms for neural networks. Vol. 39, pp. 453–474, 1992.

[28] F. Rosenblatt. The perceptron: A probabilistic model for information stage and organization in the brain. *Psychological Rebiew*, Vol. 65, No. 6, pp. 386–408, 1958.

[29] R.Reed. Pruning algorithms – a survey. *IEEE Trans. Neural Networks*, Vol. 4, No. 5, pp. 740–747, 1993.

[30] A.P. Russo. Neural networks for sonar signal processing. *IEEE Conference on Neural Networks for Ocean Engineering*, 1991.

[31] S.Amari. A theory of adaptive pattern classifiers. *IEEE Trans.*, Vol. EC-16, No. 3, pp. 299–307, 1986.

[32] S.A.Solla, E.Levin, and M.Fleisher. Accelerated learning in layered neural networks. *Complex Syst.*, Vol. 2, pp. 625–640, 1988.

[33] E. Watanabe and H. Shimizu. Algorithm for pruning hidden units in layered neural network for binary pattern classification problem. *Proc. IJCNN*, pp. 327–330, 1993.

[34] S. Yasui. A new method to remove redundant connections in backpropagation neural networks: Introduction of "parametric lateral inhibition fields,". pp. II 360–367, 1992.

[35] スメール, ハーシュ（田村一郎, 水谷忠良, 新井紀久子訳）. 力学系入門. 岩波書店, 1976.

[36] 和田安弘, 川人光夫. 新しい情報量基準 cross validation による汎化能力の推定. 電子情報通信学会誌論文誌, Vol. J74-D-II, No. 7, pp. 955–965, 1991.

[37] 郷原一寿, 内川嘉樹. 階層型ニューラルネットワークにおける学習曲面の解析. 電子情報通信学会技術研究報告, Vol. NC 90-43, , 1990.

[38] 下平丕作士. ニューラルネットワークにおける誤差逆伝播法の学習性能向上のための重み初期値の設定方法. 情報処理学会論文誌, Vol. 35, No. 10, pp. 2046–2053, 1994.

[39] 渡辺嘉二郎, 芹沢一雅, 候麗雅. 学習係数の適応調整によるニューラルネットワークの学習の高速化. 計測自動制御学会論文集, Vol. 30, pp. 1093–1099, 1994.

[40] 賈棋, 戸田尚宏, 臼井支朗. ニューラルネットワークにおける逆伝搬学習アルゴリズムの初期値設定に関する一考察. 電子情報通信学会論文誌, Vol. J73-D-II, No. 8, pp. 1179–1185, 1990.

[41] 三谷光昭, 木津徳仁, 大堀隆文, 渡辺一央. パターン認識用ニューラルネット学習高速化手法. 電子情報通信学会論文誌, Vol. J77-D-II, No. 1, pp. 211–218, 1994.

[42] 丹康雄, 加藤喜永, 江島俊. 誤差評価関数による pdp モデルの高速化. 電子情報通信学会論文誌, Vol. J73-D-II, No. 12, pp. 2022–2028, 1987.

[43] 白土浩, 五反田博, 井上勝裕, 熊丸耕介. 入力集合の可分離性からみた両極型ネットの収束上の優位性. システム制御情報学会論文誌, Vol. 11, No. 4, pp. 190–197, 1998.

[44] 甘利俊一. パターン認識の理論. 計測と制御, Vol. 7, No. 3, pp. 180–189, 1968.

[45] 萩原将文. 淘汰機能を有するバックプロパゲーション. 電子情報通信学会論文誌, Vol. J74-D-II, No. 6, pp. 812–818, 1991.

[46] 栗田多喜夫. 情報量規準による3層ニューラルネットの隠れ層のユニット数の決定法. 電子情報通信学会論文誌, Vol. J73-D-II, No. 11, pp. 1872–1878, 1990.

[47] 市村直幸, 竹内倶佳, 永井豊. パターン認識のためのクラスタ分析に基づく3層ニューラルネットワークの重み初期値の一設定法. 電子情報通信学会論文誌, Vol. J77-D-II, No. 2, pp. 301–310, 1994.

[48] 五反田博, 植田吉祥, 白土浩. ユニット極性からみた収束挙動の等価なネットワーク構成. 電子情報通信学会技術研究報告, Vol. NC 94-96, pp. 159–166, 1995.

[49] 五反田博, 植田吉祥, 川崎武士. ユニット極性の逆伝搬学習に及ぼす影響. 電子情報通信学会論文誌, Vol. J78-D-II, No. 9, pp. 1372–1382, 1995.

[50] 五反田博, 白土浩, 井上勝裕, 熊丸耕介. 多層ネットの解濃度とハードウェア化におけるスケーリング法に関する考察. 情報処理学会論文誌論文誌, Vol. 37, No. 8, pp. 1535–1542, 1996.

[51] 五反田博, 白土浩, 井上勝裕, 熊丸耕介. ネット極性による解領域の相違. 電子情報通信学会論文誌, Vol. J80-D-II, No. 2, pp. 696–699, 1997.

[52] 松永豊, 村瀬一之, 山川修, 谷藤学. 競合作用により冗長中間層素子を自動淘汰する誤差逆伝搬アルゴリズム. 電子情報通信学会論文誌, Vol. J79-D-II, No. 3, pp. 403–412, 1996.

[53] 横井邦夫, 酒井征直, 奥田利信, 郷原一寿, 内川嘉樹. 記憶面をもとにした階層型ニューラルネットワークの学習過程の解析および学習の高速化手法の提案. 信学論, Vol. J79-D-II, No. 6, pp. 1128–1133, 1996.

[54] 堀川洋. 誤差逆伝搬学習における局所解の吸引域の形状について（xor 問題の場合）. 信学論, Vol. J76-D-II, No. 10, pp. 2247–2248, 1993.

[55] 久間和生, 中山高. ニューロコンピュータ工学. 工業調査会, 1992.