1　はじめに

階層型ニューラルネットワークの応用が進む一方でネットワークの構造化が重要な課題となっている。

ネットワークの素子数や重みパラメータの初期値によって決定されるネットワークの初期構造は学習の収束性や, 未学習データに対する汎化能力などに大きく影響を与える。例えば学習対象とする関数が未知の場合, 従来の BackPropagation 学習法 (BP 法) では試行錯誤的に適当な数の隠れ素子を設定しているが, 素子数が不足すると学習が収束しない。一方、学習サンプル数に対して素子数が多すぎて自由パラメータ数が過剰になると, 未学習サンプルに対する汎化能力が低下することが理論的に明らかになっている。また実際的にも学習パラメータが増加することによる学習速度の低下や, 過学習 (オーバーフィッテイング) による汎化能力の低下が問題になる。こうしたことから学習の収束を保証し, かつ汎化能力を持たせるために適切にネットワーク構造を設定する必要がある。

学習の対象をある関数として, これが属するクラス (概念関数クラスとよぶ) に対して, ネットワークによって実現できる関数のクラス (仮説関数クラスとよぶ) がどれだけ近づけるかは学習可能性を考える上で重要な問題である。この仮説関数クラスとしてどんな関数属をとりうるかということはネットワークの構造によって決定される。ニューラルネットワークを非線形関数を基底とする関数展開として見ることで, この基底による系の性質によって仮説関数クラスを特徴づけることができる。

一方, 非線形関数の展開としてウェーブレットを用いたウェーブレット解析が注目されている。ウェーブレットによって張られる関数空間は解像度によって階層化され, 全体として多重解像度解析を形成している。また解像度を変化することで関数クラスの complexity, 近似精度を制御できる。シグモイドの組合せによって基本ウェーブレットが作れるので, ニューラルネットワーク上でウェーブレット展開を構成することは隠れ素子パラメータに制約条件を導入することで等価に実現できる。この系をネットワークに導入することで設定した解像度に応じた近似精度を与えることができる。

本研究ではウェーブレット展開を構成する隠れ素子パラメータの設定法をニューラルネットワークにおける多重解像度構造として導入し, その特長について述べる。またこれをニューラルネットワークの構造決定, 初期設定に応用し, 学習する方法を示す。さらにこの方法を用いた場合にニューラルネットワークの学習能力が改善されることを示す。

2　準備

2.1　階層型ニューラルネットワーク (MLP)

ここではフィードフォワード結合を持つ階層型ニューラルネットワークを考える。

入力素子, 隠れ素子, 出力素子の数をそれぞれ 1, m, 1 個とし, 入力から隠れ素子 $i(i = 1, \cdots, m)$ への重みを w_i, 隠れ素子 i のしきい値をθ_iとする。また隠れ素子 i から出力素子へのパラメータを v_i, 出力素子のしきい値を ϑ とし, 各隠れ素子の出力を $h_i(x)$ とすると 3 層ニューラルネットワークの出力 $g(x)$ は次の様にかける。

$$g(x) = \sum_{i=1}^{m} v_i \cdot h_i(x) + \vartheta \tag{2.1}$$

$$h_i(x) = \sigma(w_i x + \theta_i) \tag{2.2}$$

ただしシグモイド関数$\sigma(x) = 1/(1+exp(-x))$ とする。

すなわち $g(x)$ は基底関数列 $\{1, h_1, \cdots, h_m\}$ の線形和と見ることができ, これは関数系 $\{h_i\}$ に関して重み v_i を係数とした級数展開とみなすことができる。

2.2　サンプルからの学習

ある関数クラスを Fとして未知の関数 $f \in F$によって発生される有限個のサンプル集合 $\{(x_i, f(x_i))\}$ からこの関数 fを学習する問題を考える。すなわち学習モデルによって表しうる仮説関数空間 G の中を探索し, 次式を満たす関数 $g \in G$ を推定する。($d(f, g)$ は関数空間上の距離を表すものとする)

$$d(f, g) < \varepsilon$$

学習の評価を与える基準 $d(f, g)$ としては入力空間 I のもとで損失

$$R(w) = \int_{x \in I} \{f(x) - g(x)\}^2 dx \tag{2.3}$$

を考えるべきであるがたかだか有限の k個のサンプルのみから学習することを考えた場合には観測できるのは (2.3) ではなく経験損失 (2.4) になる。

$$d(f, g) = R_{emp}(w) = \frac{1}{k} \sum_{i=1}^{k} \{f(x_i) - g(x_i)\}^2 \tag{2.4}$$

学習プロセスは次のようになる。

- 概念クラス Fに対して, 学習モデルとして Fを十分近似できるような仮説関数クラス G を選ぶ。
- 仮説関数 $g \in G$ を探索するアルゴリズムによって $d(f, g)$ を最も小さくできる仮説 g^*を採択する

2.3 学習パラメータの性質

MLP の場合, 仮説関数 gは式 (2.1,2.2) のようにパラメタライズされているから, 仮説空間内の探索は実際にはパラメータ $\{v_i, \vartheta, w_i, \theta_i\}$ の探索になる。この各パラメータを説明のためベクトルによって表すことにする。

$$\boldsymbol{v} = (\vartheta, v_1, \cdots, v_m)$$

$$\boldsymbol{w}_i = (\theta_i, w_i)$$

また, 各 w_iを集合として次のように表記する。

$$W = \{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_m\}$$

式 (2.1) のように$\boldsymbol{v}$に関して gは線形である。したがってあるサンプル集合 $S = \{(x_i, f(x_i))|i = 1, \cdots\}$ が与えられたときこれにたいして $d(f, g)$ を最小にする$\boldsymbol{v}$の最適値$\boldsymbol{v}^*$はある Wについて一意に存在し, またこれを高速に計算するアルゴリズムも存在する。したがって$\boldsymbol{v}$についてはある Wにたいして必ず$\boldsymbol{v}^*$が対応しているものとして考える。

一方,fに対する Wの最適解を求めることは不良設定問題であり, 一意に定めることはできない。

式 (2.2) をあらためて

$$h_i(x) = h(x; \boldsymbol{w}_i) \tag{2.5}$$

と書き直すと, このネットワーク gによって実現できる仮説関数クラス G は非線形関数の系 $\{h(x; \boldsymbol{w}_i)|\boldsymbol{w}_i \in W\}$ によって張られる関数空間となる。このとき Wがネットワークの構造を表すことになる。

前述の通り$\boldsymbol{v}$はある Wについて一意に定まるのであるから, 学習の本質は W, すなわちネットワークの構造の探索と考えてよい。

2.4 BP 法による構造の形成

BP 法では適当な数の素子と初期値から, 誤差の最急降下法によるパラメータチューニングによってある程度学習データを反映したネットワーク構造が形成される。しかしその収束は一意でなく, 素子数や初期値への依存性が大きい。

BP による構造の学習過程を説明するために BP 法を抽象化して表すと次のように書ける。

1. 初期構造 $W^{t=0}$を設定する。
2. W^tについて $d(f, g)$ を最小化する$\boldsymbol{v}^t$を求める。
   ($d < \varepsilon$なら終了。)
3. $\boldsymbol{v}^t$を固定してさらに $d(f, g)$ がより減少するように W^{t+1}を定める。
4. 2. へ戻る。

実際に構造 Wの修正は上の 3. で行われているが,$\boldsymbol{v}^t$にとっての W^{t+1}は必ずしも意味のある構造とは限らない。また学習自体は決定的に行われ, その収束状態は初期状態 W^0に強く依存することがわかる。

また一般に用いられる方法では W^0として小さな乱数を与えることが多い。この場合, 素子数が多くなると各 $\boldsymbol{w}_i$は原点近くに密集することになるため, Wの探索効率は非常に悪い。実験的にも互いに線形従属で, 冗長な基底が発生することが確認された [9]。

したがって従来の BP 法による学習では最適な構造が常に形成されるとは限らず, 実際には与えられた素子数や初期値に応じて (最適とは限らない) 構造ができている。

より適した構造を獲得するためには, 積極的に構造に関する評価を入れ, それにより望ましい構造をネットワークに形成させるといったアプローチが必要であるように思われる。

3 多重解像度構造

探索空間 $w - \theta$の中でどのように構造 Wを探索すべきかについて考える。

例えば各基底関数 $h(x; \boldsymbol{w}_i)$ が正規直交系を成すように Wが定まる場合には古典的関数近似理論による議論が成立する。しかし, ニューラルネットの場合, 一般的に各基底が全て直交するようにはならない。しかし, 有限個の非直交基底で概念関数クラスを十分近似できるような仮説関数クラスを構成することは可能である。

3.1 離散ウェーブレット展開

非直交ウェーブレットを用いて任意の $f \in L^2(\boldsymbol{R})$ を展開する方法がある。$\psi(x)$ をアドミッシブル条件を満たす基本ウェーブレット関数として次のような系を構成する。[1](ただし $m, n \in \boldsymbol{Z}$)

$$\psi_{m,n}(x) = \psi(a_0^m x + nb_0) \tag{3.6}$$

この系によって次のような関数展開を考えることができこれを離散ウェーブレット展開という。

$$\sum_{m,n} c_{m,n}\psi_{m,n}(x) \tag{3.7}$$

[1] ここでは説明のため [1] とは m の符号を逆にしている。

ウェーブレット展開の特徴はこれが多重解像度解析になることである。m は解像度レベルを表すパラメータになり,n はその中で基底のシフトパラメータである。n の範囲については基底が入力空間を覆うように定めることができる。ここに解像度 m を $m_{min}, m_{min+1}, \cdots, m_{max}$ のように多重に設定することになる。

これが任意の $f \in L^2(\boldsymbol{R})$ を近似できるような基底ψと係数 (a, b) の選び方がある。このような条件を満たす ψ, a_0, b_0, について調べられている (例えば [1])。またシグモイド関数σを組み合わせてψを作った場合にも, この条件を満すことができる [5]。

ここで次のように 3 個のシグモイド関数で基本ウェーブレットを作る。

$$\psi(x) = \sigma(x-k) - 2\sigma(x) + \sigma(x+k) \tag{3.8}$$

式 (3.6) において $a_0 = 2$,$b_0 = k$とすると, このウェーブレット展開は次のようにかけ,

$$g(x) \quad = \quad \sum_{m,n} c_{m,n}\psi(2^m x + nk) \tag{3.9}$$

これに (3.8) を代入して係数をあらためて

$$c'_{m,n} = c_{m,n-1} - 2c_{m,n} + c_{m,n+1}$$

とするとこのウェーブレット展開は次のニューラルネットで実現できる。

$$g(x) \quad = \quad \sum_{m,n} c'_{m,n}\sigma(2^m x + nk) \tag{3.10}$$

すなわち上式 (3.10) を満たすようにパラメータを次のように設定すればよい。

$$\boldsymbol{w}_{m,n} = (w_{m,n}, \theta_{m,n}) = (2^m, nk) \tag{3.11}$$

これがニューラルネットワークをウェーブレット展開とするための条件になる。

(このときψを構成するのにσを 3 個使うが, 全体として系を構成した時,σが重複するように b_0 を選べば[2], 係数のみ考慮することで必要なσの数は減少でき,ψの 3 倍にはならない。)

3.2 多重解像度構造の導入

ウェーブレット展開を構成するように (2.1),(3.11) にしたがって隠れ層のパラメータ w, θを定めた時, これをニューラルネットワークにおける多重解像度構造とよぶ。

[2] $b_0 = k, 2k$などとする

$$g(x) \quad = \quad \sum_{m,n} v_{m,n}\sigma(2^m x + nk) + \vartheta \tag{3.12}$$

多重解像度構造をとることで m, n により Wが決定され, $v_{m,n}$については前述のように最適値を求めることができる。このように考えることで探索すべきパラメータが $w, \theta \in \boldsymbol{R}$から $m, n \in \boldsymbol{Z}$になるから探索空間を強く制約できる。n の範囲については入力空間 Iによるθの制約条件 (付録参照) を用いることができ, 任意性のあるパラメータは解像度 m だけになる。$m_{min} \leq m \leq m_{max}$ とするとニューラルネットの場合, 基底としてしきい値 ϑが含まれているからこれを最低解像度 $m_{min} = -\infty$ と近似する。m については m_{max}が増大するにつれ解像度が高くなり近似精度が向上するという性質がある。しかし, サンプル数が少ない場合には過度の解像度で展開することは出来ない。このとき $w = 2^m$であるから解像度が高くなるとともに wが増大することになり,wが極端に増大することで過学習が起きることと関連づけることができる。したがって m_{max}については要求する近似精度とサンプル数に注意して設定しなければならない。

$m_{max} = M$とした多重解像度構造によって表される仮説関数クラスを G_Mとかくことにするとこのような関数空間は階層化されていて,

$$G_1 \subset G_2 \subset \cdots \subset G_\infty$$

が成立している。すなわち通常の MLP と異なり解像度の概念によって, 学習モデルの階層化が厳密に成立している。これは AIC や Structual Risk Minimization Theory [3, 4] を適用する場合に非常に重要な性質である。

多重解像度構造を構成するように隠れ層を構成した場合には $f \in L^2(\boldsymbol{R})$ にたいして解像度に応じた近似能力を持つ系 $W = \{(w_{m,n}, \theta_{m,n})\}$ が設定できるから, 学習は Wにたいする出力素子の最適パラメータ$\boldsymbol{v}$を求めることで完了する。

ところで多重解像度構造をとる gは任意の $f \in L^2(\boldsymbol{R})$ を近似する能力をもっているのである特定の fを学習する場合には, 全ての基底$\psi_{m,n}$ を必要とすることはあまりない。このとき不必要な基底を判別する際にウェーブレット係数 $c_{m,n}$ を用いて$\psi_{m,n}$の有効性を評価することができる。さらに従来のσによる基底表現では, 追加, 削除を行うとそれまでの学習結果が無駄になることがある。基本となる基底をσでなくψとすると, これはコンパクトサポートで, かつ平均$\int \psi dx$ が 0 であるため, 基底の追加, 削除を行った場合にそれまでの学習結果に与える影響が少ないと考えられる。こうしたことから多重解像度構造をとることによって素子の追加, 削除が容易に行える。

以上の様なメリットが考えられるが, 今回は特に初期構造としてウェーブレット展開を導入した場合の学習能力の改善について実験的に評価する。

4 多重解像度構造による学習能力の改善

ここで多重解像度構造を初期構造として考える。さらに適切な初期構造の設定のため次のような学習アルゴリズムを考える。

学習アルゴリズム (MR)

1. 解像度 m_{max} を定め, 入力空間 $I = [x_{min} : x_{max}]$ から w, θ の範囲を求める。$((3.11)$, 付録参照$)$
2. その範囲の中で (3.12) により隠れ層を構成する。
3. $\boldsymbol{v}$ の最適値を求める。
4. 全てのパラメータ $(\boldsymbol{v}, \boldsymbol{w})$ について BP 法によりチューニングする。

この学習アルゴリズム (MR) による学習能力をシミュレーションによって確認する。またこれを同数の素子を持つ通常の BP 法 (normal) [3] と比較する。

[学習条件]

- 学習係数 0.05, 慣性項 0.02, サンプル数 50
- サンプルは $I = [-1 : 1]$ から一様ランダムに発生する。
- 次の sin 関数の族 $s_i(x)$ を学習する。

$$s_k(x) = sin(k\pi x)$$

MR の 3. で $\boldsymbol{v}$ を求めるのに,BP 法と同条件の最急降下法を用いることで学習能力を直接比較できるようにした。具体的には出力層のみ,BP 法と同じ学習係数で 500 回学習する。したがって MR と Normal は初期設定以外, 全て同条件である。

実験結果を学習曲線で表す。解像度レベルは $1 \leq m \leq M$ とし,$M = 2, 3$ の場合を示す。縦軸は平均二乗誤差 $(\mathrm{R.M.S.E})$, 横軸は学習回数である。

ここで 4. で隠れ素子パラメータにたいして BP 法を用いなければウェーブレット展開と等価であるが, ウェーブレット展開は f についてシフト不変でなく, シフト変化にたいして敏感である。すなわちある関数について良好な近似精度が得られる解像度であったとしても, わずかにシフトしていると, 同じ解像度でも精度が悪化する

[3] 初期値を [-0.3:0.3] の一様乱数によって定める。

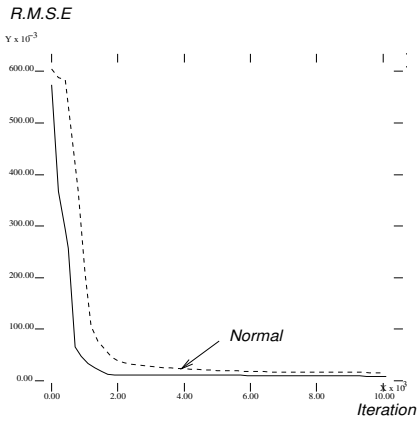

図 1: 学習曲線:$s_2(x), M = 2$

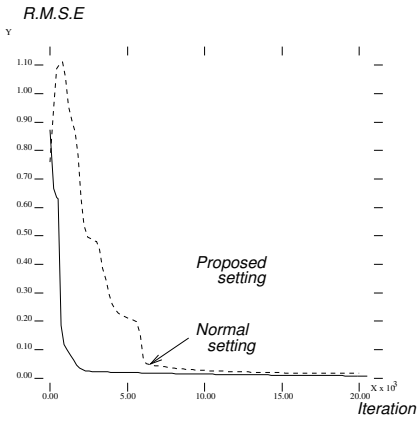

図 2: 学習曲線:$s_3(x), M = 2$

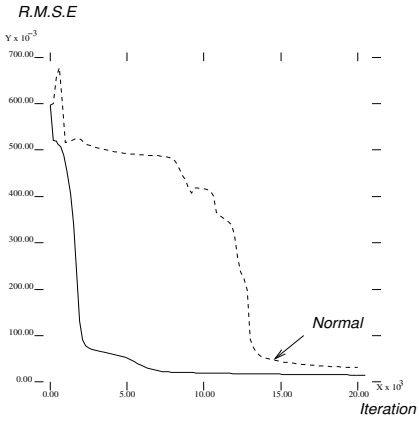

図 3: 学習曲線:$s_4(x), M = 2$

ことが起こりうる。以上の理由から多重解像度構造を初期値とし, さらに BP 法によってパラメータチューニングを行っている。

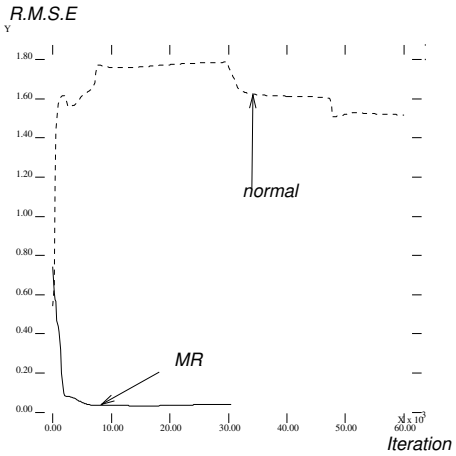

図 4: 学習曲線:$s_5(x), M = 3$

5 考察

学習対象の関数 $s_k(x)$ の kが大きい程, 関数としての complexity が大きい。図 1,2,3からわかるように kが小さい場合には Normal と MR の差は大きくないが,kが大きくなるにつれ Normal よりも MR の方が速くなっている。つまり MR の初期設定により複雑な関数ほど学習の高速化に効果がある。また図 4,5 の $s_5(x), s_6(x)$ では Normal は今回の条件では学習が収束しなかったが MR は学習が進んでいる。これは complexity にたいして, サンプル数が 50 と, 比較的少ないことにも起因している。(サンプル数を多く与えた場合には Normal でも収束が見られることがあった。) この点で MR はサンプル数が比較的少ない条件のもとでも学習能力が改善できるといえる。

6 まとめ

離散ウェーブレット展開を実現するニューラルネットワークの多重解像度構造を提案し, これを学習の初期構造の設定に応用した。またシミュレーションにより学習能力が従来の方法と比較して, 学習速度, 収束性の点で改善されることを示した。

今後の課題としてはノイズを含んだサンプルから学習する場合や, 汎化性を考慮する場合における学習能力の評価がある。またウェーブレットはテンソル積を考えることで入力が多次元の場合に拡張できるが, 次元が高くなるにつれ, 必要とする基底が膨大になる。通常のニューラルネットワークを多次元の問題に適用した場合, これよりもはるかに少ない素子数で済むことが多く, こうした問題についてはさらに経済的な構造化や, プルーニングアルゴリズムを考える必要がある。

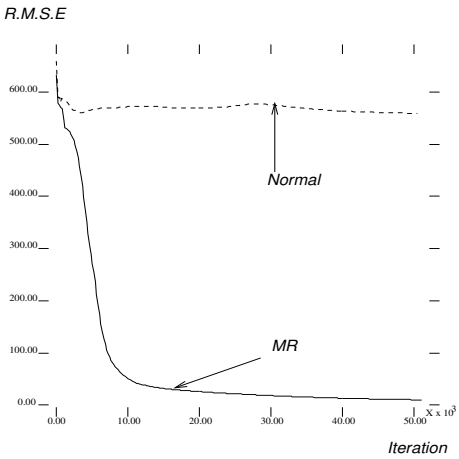

図 5: 学習曲線:$s_6(x), M = 3$

謝辞

研究の機会を与えていただいた諏訪基情報科学部長に感謝いたします。また執筆にあたり御指導頂いた梅山伸二情報数理研究室長, 麻生英樹主任研究官, ウェーブレットの輪講を企画していただいた田中勝研究官, ならびに日頃より御討論いただく栗田多喜夫主任研究官をはじめとする情報数理研究室の皆様に深く感謝いたします。

参考文献

[1] I.Daubechies:"Ten Lectures on Wavelets" ,CBMS ,SIAM (1992).

[2] S.G.Mallat:"A Theory for Multiresolution Signal Decomposition: The Wavelet Representation",IEEE Trans. on PAMI,vol.II.No.7,july(1989).

[3] V.N.Vapnik:"Estimation of Dependencies Based on Empirical Data",Springer-Verlag,NewYork(1982).

[4] V.N.Vapnik:"Principles of Risk Minimization for Learning Theory",Advanced in Neural Information Processing System 4,Morgan Kaufmann,(1992).

[5] Y.C.Pati,P.S.Krishnaprasad:"Analysis and Synthesis of Feedforward Neural Netwaorks Using Discrete Affine Wavelet Transformations",IEEE Trans. on Neural Networks vol4,No.1,January(1993).

[6] 佐藤:" ウェーブレット理論の数学的基礎", 日本音響学会 vol.47 No.6 ,pp.405-423(1991).

[7] 麻生 :"学習の数理的理論の展開", 電子技術総合研究所彙報, vol.56, pp.602-635 (1992).

[8] 栗田:" 階層型ニューラルネットのパラメータの最尤推定について", 信学技報 NC91-36(1991).

[9] 本村, 高橋, 富田:"MLP の構造学習のための隠れ素子の解析と構成法", 信学技報 NC92-6(1992).

付録

パラメータの制約条件

構造 W の探索空間は無限であるが, 各基底を表す $\boldsymbol{w}_i$ についての性質は調べることができる [9]。

まず $x \in I$ が次のようであると仮定する。

$$x_{min} < x < x_{max}$$

つぎにシグモイド関数 $\sigma(x)$ を考えると,x が 0 から増大または減少するにつれ $d\sigma(x)/dx$ は単調に減少し 0 に漸近する。これが十分漸近したときの $|x|$ を $L(>0)$ とすると,$\sigma(x-x_0)$ の変動が無視できるようになるのは,$|x-x_0| \geq L$ であるから

$$x_0 \leq x_{min} - L$$

または

$$x_0 \geq x_{max} + L$$

となる場合に,$\sigma(x-x_0)$ は $x \in I$ において常にほぼ定数値をとることになる。したがってこれをさけるには次式を満たせばよい。

$$x_{min} - L < x_0 < x_{max} + L$$

同様に $h(x;w,\theta) = \sigma(wx+\theta) = \sigma(w(x-\frac{\theta}{w}))$ について見ると L の代わりに $\frac{L}{|w|}$,$x_0 = -\frac{\theta}{w}$ を考えればよいから,

$$x_{min} - \frac{L}{|w|} < -\frac{\theta}{w} < x_{max} + \frac{L}{|w|}$$

これを満たさない基底関数 h は $x \in I$ の中で展開 (2.1) を考えたときには冗長であり, 無視できる。したがって $h(x;w,\theta)$ として考えなければいけないパラメータ w,θ は上式を変形して

$$-x_{max}w - L\frac{w}{|w|} < \theta < -x_{min}w + L\frac{w}{|w|} \qquad (0.13)$$

を満たすものだけである。

また $\sigma(x)$ の点対称性からパラメータ空間には対称性が存在している。W の中で任意の $\boldsymbol{w} = (w,\theta)$ を符号を逆転した $-\boldsymbol{w} = (-w,-\theta)$ で置き換えた W' を考えると必ず

$$g(x;W,\boldsymbol{v}^*) = g(x;W',\boldsymbol{v}'^*)$$

を満たすような $\boldsymbol{v}^*, \boldsymbol{v}'^*$ が存在する。つまり, ネットワークの出力不変のまま $\boldsymbol{w}$ の符号反転をすることができる。この操作を有限回行って例えば $w \geq 0$, または $\theta \geq 0$ のようにすることができる。

以上のことを考慮してパラメータ空間 $(w-\theta)$ の中で探索すべき領域を限定することができる。例えば $-X < x < X$,$w \geq 0$ としたとき (0.13) から次式が成立する。

$$-Xw - L < \theta < Xw + L, (w \geq 0)$$

これを図示すると図 6のようになる。

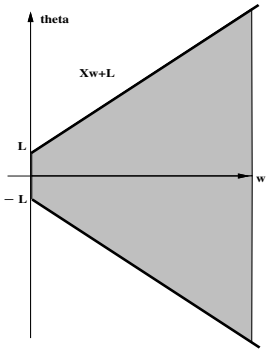

図 6: パラメータ空間 $(w-\theta)$