ニューラルネットの基礎数理（2）

上坂　吉則

3. 学習の問題点

前節で紹介した誤差逆伝搬法は確かに強力な学習法ではあるが，3つの大きな問題をかかえている．

この学習は本質的には最急降下法であるから，一般には誤差Eの極小点が求まってしまう．しかし，欲しいのは最小点であるから，重みの初期値をどのように選べばよいかを考えなければならない．これが第1の課題である．しかし，理論的には現在のところお手上げであり，多くの初期値を試みてその中から相対的に最良なものを選んで我慢するしかない．

第2の問題点は学習が達成されるまでの，つまり，誤差の極小点に到達するまでの時間が，時には異常に長くかかるという点である．この難点については，いわゆる数値解析の観点から現在さまざまな形で研究されている（たとえば，文献［1，17］参照）．

誤差Eは重みw_{ij}, v_jの無限回連続微分可能な多変数の関数であるから，これを一般に$\mathbf{R}^n$から$\mathbf{R}$への必要な回数までの連続微分可能な関数とし，その変数をxで表わすことにしよう．いま，Eが$x^* \in \mathbf{R}^n$で極小になっているとすると，Eはこの点の近傍で2次関数によって

$$E(x)=\frac{1}{2}(x-x^*)^t A(x-x^*)+E(x^*) \tag{3.1}$$

と近似できる．ここに$A=[a_{ij}]$はn次の対称な正定値行列である．そこで

$$\frac{\partial E}{\partial x_i}=\sum_{j=1}^{n} a_{ij}(x_j-x_j^*) \tag{3.2}$$

に注意して，この関数に前節の誤差逆伝搬法を用いると，学習の漸化式は$k=0, 1, 2, 3\cdots\cdots$に対して

$$x(k+1)=x(k)-\Delta A(x(k)-x^*) \tag{3.3}$$

となる．この漸化式によって$x(k)$が極小点x^*に近づく速さを表わすのに，**通常誤差の減衰率**：

$$\varepsilon=\inf_{\Delta} \limsup_{k\to\infty} \frac{\|x(k+1)-x^*\|}{\|x(k)-x^*\|} \tag{3.4}$$

が用いられる．ここに$\|x\|$はベクトルxのユークリッドノルムである．

式(3.3)の行列Aが対角化でき，その固有値がすべて正であることに注意すると，上の減衰率はAの最大および最小固有値λ_M, λ_mを用いて

$$\varepsilon=\frac{\lambda_M-\lambda_m}{\lambda_M+\lambda_m} \tag{3.5}$$

と計算される．

一方，式（3.3）の漸化式に少し手を加えて，$k=1, \cdots$に対して

$$x(k+1)=x(k)-\Delta_1 A(x(k)-x^*)+\Delta_2(x(k)-x(k-1)) \tag{3.6}$$

と，2階の差分方程式にしてみると，このときの誤差の減衰率：

$$\varepsilon'=\inf_{\Delta_1, \Delta_2} \limsup_{k\to\infty} \frac{\|x(k+1)-x^*\|}{\|x(k)-x^*\|} \tag{3.7}$$

は

$$\varepsilon'=\frac{\sqrt{\lambda_M}-\sqrt{\lambda_m}}{\sqrt{\lambda_M}+\sqrt{\lambda_m}} \tag{3.7}$$

と計算される．

この2つの減衰率を比較すると容易にわかるように，$\varepsilon' \leq \varepsilon$が成り立っている．つまり，手を加えた漸化式による学習の方が一般に速く極小点に収束すると考えられる．このような学習速度の改良の試みが数値解析の立場からいろいろ試みられている［1，17］．

さらに，第3の問題点として“学習”機械としてのより重要な課題を考えなければならない．2節で紹介した“学習”が行なっていることは，数学的には，学習データ（Sとその上のdの値）をできるだけ満たすようにニューラルネットhで目標関数dを近似することに他ならない．しかし，本音はそれだけではなく，$X-S$の上でもdを近似したいという淡い期待をもっている．つまり，学習時に教えられていないパターンに対しても回路が正しく応答してくれること，すなわち，**学習の汎化性**を期待している．これが単なる関数近似と学習を区別する重要なポイントである．

うえさか　よしのり　東京理科大学　理工学部
〒278 野田市山崎2641

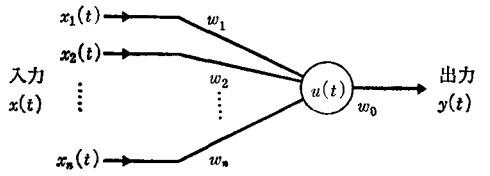

図 4.1 アナログ型動的決定論的ニューロンモデル

しかし，有限個のパターンに対する d の値が与えられただけでは，それ以外のパターンにおける d の値を知るすべは一般に存在しない．目標関数 d に関して何らかの予備知識（モデル）が必要である．われわれの学習では，“重みを種々変えて得られるすべての h の族に d が属している”と暗に仮定していると考えられる．この仮定を満たす d を相手にしているときには，d の推定はある程度成功すると期待できる．しかし，そうでないときはお手上げである．

そこで隠れ素子の数 m をうんと増やしておいてニューラルネット h の族，すなわち，モデルを初めから大きくとっておくことになる（$m\to\infty$ でほとんどすべての関数（回路）が実現できることが知られている［4］）．このときは，しかし，上で指摘したように学習（推定）がきわめて困難になる．

こうして目標関数とそれに対するモデルの規模と学習サンプルの大きさとの関連を議論することが本質的に重要な課題となってくる．層状回路についてのこの種の課題に対する本格的な研究はこれからではあるが，もっと広い枠組みのなかではすでに情報量基準（ＡＩＣ）によるシステムの構造推定［11］，万能学習機械の理論［8, 10］，学習機械の複雑さと学習可能性の関係［2, 3, 15］，学習可能性の一般論［12, 13］，計算論的学習可能性の理論［5, 19］など展開されている．

4. 決定論的最小値探索機械

時間 t の実数値関数 $x_1, \cdots\cdots, x_n$ を入力したとき，内部電位 u と出力 y が次の微分方程式と関数 tanh：

$$(4.1)\qquad \frac{du}{dt}=-\frac{u}{\tau}+\sum_{i=1}^{n}w_ix_i+w_0,\quad y=\tanh u$$

にしたがうようなニューロンモデルを考える（図**4.1**）．ここに u はニューロンの内部電位と呼ばれる時間の関数，w_i や w_0 は重みやしきい値（の符号を反転したもの）と呼ばれる定数，τ は時定数と呼ばれる正の定数である．この微分方程式からわかるように，重み w_i が正（負）のときは入力 x_i の正負に応じて内部電位 u は上昇または下降（下降または上昇）する傾向があり，この

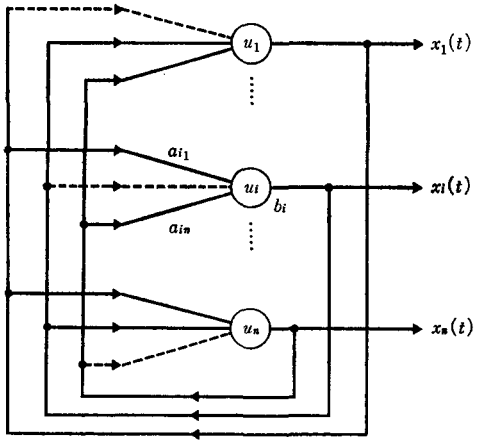

図 4.2 決定論的最小値探索機械を実現するフィードバック型ニューラルネット

ことは過去の入力の状況にも依存する．この意味で，このニューロンモデルは記憶をもっているということもできる．また，出力 y は内部電位が非線形に変換されて生じ，その値は開区間（-1, $+1$）の値をとる．このような情報処理素子を**アナログ型動的決定論的モデル**という．

いま，このようなニューロン素子を n 個用意し，その各出力をすべての素子の入力にフィードバックすることによって得られる回路，すなわち，相互結合型の回路を考える（**図4.2**）．そうするとこのニューラルネットの動作は次のような力学系（連立の微分方程式系）：

$$(4.2)\qquad \frac{du_i}{dt}=-\frac{u_i}{\tau}+\sum_{j=1}^{n}a_{ij}x_j+b_i\ (i=1,\cdots\cdots,n),$$

$$(4.3)\qquad x_i=\tanh u_i\ (i=1,\cdots\cdots,n)$$

で表わされることになる［7］．ここに a_{ij} は i 番目の素子の j 番目の入力に対する重みであり，b_i は i 番目の素子のしきい値（の符号を反転したもの）であり，時定数 τ は共通の値をとることにしている．

ここで n 次元ユークリッド空間 $\mathbf{R}^n$ で定義された実数値関数：

$$(4.4)\qquad E(x_1,\cdots\cdots,x_n)=-\frac{1}{2}\sum_{i,j=1}^{n}a_{ij}x_ix_j-\sum_{i=1}^{n}b_ix_i+\frac{1}{\tau}\sum_{i=1}^{n}\int_0^{x_i}\tanh^{-1}(x)dx$$

を用意し，各ニューロンの重み係数に関して

$$(4.5)\qquad a_{ii}=0,\ i\neq j\Rightarrow a_{ij}=a_{ji}$$

を仮定して，E を x_i で偏微分してみると

$$(4.6)\qquad \frac{\partial E}{\partial x_i}=-\sum_{j=1}^{n}a_{ij}x_j-b_i+\frac{1}{\tau}u_i$$

が得られる．したがって $x_1, \cdots\cdots, x_n$ が上の微分方程式

の解ならば，式 (4.2) から明らかなように，$-\partial E/\partial x_i = du_i/dt$ となり，また，Eは $x_1, \cdots\cdots, x_n$ を通して時間の関数となる．このEを時間で微分してみると

$$(4.7) \qquad \frac{dE}{dt} = -\sum_{i=1}^{n}(1-x_i^2)\left(\frac{\partial E}{\partial x_i}\right)^2 \leq 0$$

と計算されることがわかる．したがって関数Eは力学系の軌道上で時間の進行に伴って非増加であり，式 (4.7) の等号が成立するのはEの極小点においてである．

いま，τを十分大きくとっておくと，Eの多項式部分が超立方体 $[-1, +1]^n$ 上で極小値をとる点，すなわち，この立方体の頂点の近くで上の力学系の状態は停留することになる．したがって集合：

$$(4.8) \qquad X=\{x|x=(x_1, \cdots\cdots, x_n), x_i=-1, +1\}$$

で定義された関数，すなわち，2値をとる変数の2次関数：

$$(4.9) \qquad F(x_1, \cdots\cdots, x_n) = -\frac{1}{2}\sum_{i,j=1}^{n} a_{ij}x_ix_j - \sum_{i=1}^{n} b_ix_i$$

の極小値（の正確な定義は後述）あるいは最小値がこの力学系の平衡点として求められることが推察される．これがニューラルネットによる**最小値探索機械**の基本的な仕組みである．

これまでの議論からわかるように，τが有限の値をとっている限り，最小値ないしは極小値が得られる可能性を厳密に保証するのは難しい．そこで以下では式 (4.2) においてτを無限大にした極限での理論から探索の可能性を正確に見ることにしよう．なお，τを無限大にするということは，式 (4.1) からわかるように内部電位 u_i が正負の無限大になり得るわけで，その電位に耐えるような理想的なニューロン素子から成るニューラルネットの力学系を考察することに相当する．事態をこのように理想化することによって，後にわかるように，事の本質が厳密な形で見えてくるのである [14, 16].

それではこれから扱う最小値問題を整理しておくことにする．巡回セールスマン問題やnクイーン問題など多くの組合せ的最適化問題は2値をとる多変数の実数値関数の制約付き最小値問題に帰着させることができる [16]．そこで次のような最小値問題を考えることから議論を始めることにしよう．

問題 4.1 FをX上で定義された2次関数（式 (4.9) 参照）とし，SをXの部分集合とする．このとき，S上でのFの最小値と最小点：

$$(4.10) \qquad F_{\min} = \min\{F(x)|x\in S\},$$

$$(4.11) \qquad x_{\min} = \arg\min\{F(x)|x\in S\}$$

を求めよ．

はじめに，次の定理が示すように，多くの場合この制約を容易に取り外すことができることを注意しておこう．

定理 4.1 Sから定まる2次関数$G: X\to\mathbf{R}$で

$$(4.12) \qquad G(x)\begin{cases} =0, & x\in S \text{ のとき}, \\ >0, & x\notin S \text{ のとき} \end{cases}$$

を満たすものが存在するとする．このとき，正の定数cを用いて

$$(4.13) \qquad H(x)=F(x)+cG(x)$$

とおくと

$$(4.14) \qquad \begin{aligned} &x_{\min} = \arg\min\{H(x)|x\in X\} \\ \Rightarrow\ &x_{\min} = \arg\min\{F(x)|x\in S\} \end{aligned}$$

が成り立つような定数cが存在する．

次に，目的関数Fの変数 x_i は ± 1 しかとらないことに注意すると，係数 a_{ij} は，一般性を失うことなく，式 (4.5) を満たしているとしてよいことが容易に示される．

さらに，式 (4.9) から1次の項を次のようにして落とすことができる．すなわち，

$$(4.15) \qquad A=\begin{bmatrix} a_{11}\cdots\cdots a_{1n} \\ \cdots\cdots \\ a_{n1}\cdots\cdots a_{nn} \end{bmatrix}, \quad b=\begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}$$

から

$$(4.16) \qquad B=\begin{bmatrix} 0 & b^t \\ b & A \end{bmatrix}$$

なる$n+1$次の正方行列を作り，この2次形式：

$$(4.17) \qquad G(y)=-\frac{1}{2}y^tBy, \quad y=(x_0, x_1, \cdots\cdots, x_n)^t$$

を考える．このとき次の定理が成り立つことを容易に示すことができる．

定理 4.2

$$\begin{aligned} &(x_0^*, x_1^*, \cdots\cdots, x_n^*)^t \\ &=\arg\min\{G(y)|y\in\{-1, +1\}\times X\} \\ &\Rightarrow \arg\min\{F(x)|x\in X\} = x_0^*(x_1^*, \cdots\cdots, x_n^*)^t \end{aligned} \qquad (4.18)$$

以上のことから，ほぼ一般性を失うことなく次の問題を考えればよいということになる．

問題 4.2 Xから$\mathbf{R}$への関数：

$$(4.19) \qquad F(x)=-\frac{1}{2}x^tAx$$

のX上での最小値 $F_{\min}$ と最小点 $x_{\min}$ を求めよ．ここに $A=[a_{ij}]$ において，式 (4.5) が満たされているとする．

問題 4.2 を解くために，Fの定義域をn次元ユークリッド空間に拡大して得られる関数Eを用意し，このEを

用いて次の微分方程式系，すなわち，力学系：

$$(4.20)\quad \frac{du_i}{dt}=-\frac{\partial E}{\partial x_i},\quad x_i=\tanh(u_i)$$

を考える．このとき，E を**目的関数 F から導かれたエネルギー**と呼ぶ．これを Hopfield らが扱った力学系 (4.2) と較べると，減衰項 $-u_i/\tau$ が落ちているが，これはさきに述べたようにニューロン素子をある意味で理想化したことに相当する．

さて，式 (4.20) において u_i を消去して x_i だけに関する方程式を作ると

$$(4.21)\quad \frac{dx_i}{dt}=-(1-x_i^2)\frac{\partial E}{\partial x_i}=(1-x_i^2)\sum_{j=1}^{n}a_{ij}x_j$$

が得られる．いま，素子の出力 x_i の n 組 $x=(x_1,\cdots\cdots,x_n)$ に着目すると，これは上の力学系の状態を表わしており，式 (4.20) の第2の式からわかるように，n 次元立方体 $C=[-1,+1]$ の中を時間とともに移動し，さまざまな軌道を描くことになる．この微分方程式系は非線形でもあり，解析的に解くことは難しい．しかし，いわゆる微分方程式の定性的理論［6］を援用することによっていくつかの重要な性質を明らかにすることができるが，以下でその主なものをまとめておこう［14, 16］.

定理 4.3 力学系 (4.21) において時刻0で立方体 C の中から出発するとき，エネルギー E は時間に関して非増加であり，$dE/dt=0$ となるのは $i=1,\cdots\cdots,n$ に関して

$$(4.22)\quad x_i=\pm1 \text{ または } \sum_{j=1}^{n}a_{ij}x_j=0$$

のときかつこのときに限る．

定理 4.4 立方体 C の相隣り合った頂点における目的関数 F の値の差：

$$(4.23)\quad F(v_1,\cdots\cdots,v_i,\cdots\cdots,v_n)-F(v_1,\cdots\cdots,-v_i,\cdots\cdots,v_n)$$

は力学系 (4.21) のヤコビ行列の固有値に等しい．

ここで目的関数 F の極小に関する概念を明確にしておこう．立方体 C の頂点 $v=(v_1,\cdots\cdots,v_n)(v_i=\pm1)$ が F の**極小点**であるとは

$$(4.24)\quad \forall i: F(v_1,\cdots\cdots,v_n)<F(v_1,\cdots\cdots,v_{i-1},-v_i,v_{i+1},\cdots\cdots,v_n)$$

が成り立つことをいう．また，式 (4.24) において少なくとも1つの $<$ が $\leq$ になっている場合，v を**広義極小点**という．v が極小点でも広義極小点でもないとき，**非極小点**であるという．

定理 4.5 目的関数 F から導かれるエネルギー E をもつ上の力学系において

1° F の極小点は漸近安定である；

2° F の広義極小点は安定なことも不安定なこともある；

3° F の非極小点は不安定である．

定理 4.6 目的関数 F から導かれるエネルギー E をもつ上の力学系において，立方体 C の内部の点は漸近安定ではない．

いま，目的関数 F の係数行列 A が正則だとすると（そして多くの場合実際そうである），式 (4.22) によれば，立方体 C の内部には平衡点としては原点があるだけである．したがって，上の定性的性質を考慮すれば，立方体内の原点以外の任意の点を初期値として出発すれば，ほとんどの場合，その漸近安定点として目的関数 F の極小点を求めることはできる．

しかし，われわれが欲しいのは F の最小点である．極小点の中には当然最小点が存在するから，うまい初期値を選ぶことにより最小値を得る可能性はある．

以下では最小値を与えてくれる初期値をどう選んだらよいかについて検討してみよう．この問題を**初期値設定問題**というが，これがニューラルネットによる最小値探索法の唯一のしかも最大の難点である．

x^* を漸近安定点とし，この x^* に近づいていくようなすべての軌道の和集合を x^* の**たらい**と呼んでいる．この意味では，最小点のたらいの中から出発すれば，必ず最小値が得られることになる．しかしこのたらいを具体的に求めることは上の微分方程式を解くのと同程度に困難である．

そこで初期値をランダムに選んで，どの程度の割合で最小値が得られるかを実験的に調べてみることにする．$n=10$ の目的関数から導かれるエネルギーを持つ力学系を考え，目的関数の係数行列 A をランダムに選んで固定しておく．そして，初期値を集合：

$$(4.25)\quad S(d)=\{(x_1,\cdots\cdots,x_n)|\forall|x_i|\leq d\}$$

の中からランダムに選んで力学系を駆動する試行を多数回行ない，最小値が得られた割合を記録する．この値は一般に $S(d)$ の大きさ d に依存する．実際，1000回の試行を種々の d に対して行なってみると $d=0.9, 0.8, \cdots\cdots, 0.1$ に対して探索の成功率は，それぞれ 53.3%，61.4%，65.4%，71.5%，79.2%，88.6%，96.1%，99.8%，100.0% と得られる．このように d が小さくなる，すなわち，$S(d)$ が狭くなるほど高い確率で最小値が得られる．いいかえれば，最小値を与えるたらいの点が $S(d)$ に含まれる割合が，d の増加とともに多くなる

ということである．この傾向は多くの種類の目的関数に対して実際観測され，このことは次の予想を示唆しているように思われる：

予想 4.1 状態空間の原点を中心とする小立方体 $S(d)$ の中からランダムに初期値を選んだとき，目的関数の最小点が得られる確率を $P(d)$ で表わす．このとき

$$\lim_{d\to 0} P(d)=1 \tag{4.26}$$

が成り立つ．

参考文献

[1] 麻生：誤差逆伝搬学習の数理的性質，電子情報通信学会技術報告，**PRU 89-14** (1989).

[2] Baum, E. B. & Haussler, D.: What size net gives valid generalization?, *Neural Computation*, **1** (1989), 151-160.

[3] Cover, T. M.: Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition, *IEEE Trans. on Electronic Computers*, (1965), 326-334.

[4] Funahashi, K.: On the approximate realization of continuous mappings by neural networks, *Neural Networks*, **2**, 3 (1989).

[5] Gold, E. M.: Language identification in the limit, *Information and Control*, **10** (1967), 447-474.

[6] Hirsch, M. W. & Smale, S.: 力学系入門，岩波，1974.

[7] Hopfield, J. J. and Tank, D. W.: "Neural" computation of decisions in optimization problems, *Biol. Cybern.*, **52** (1985), 141-152.

[8] Kovalevsky, V. A.: Recent advances, in statistical pattern recognition, *Proc. 4-th Int'l Joint Conf. on Pattern Recognition*, (1978), 1-12.

[9] 大須賀，佐伯編：知識の獲得と学習，オーム社，1987.

[10] 尾関：万能学習機械は存在するか，数理工学研究会シンポジウム，1979-01.

[11] 坂元他：情報量統計学，共立出版，1983.

[12] Uesaka, Y. et al.: A theory of learnability, *Kybernetik*, **13** (1973), 123-131.

[13] 上坂：学習可能性と線形空間，電子通信学会論文誌，*J66-A* (1983), 12, 1151-1158.

[14] 上坂：2値変数の実数値関数から導かれるエネルギーを持つニューロン回路網の安定性について，電子通信学会技術研究報告，**PRU 88-6** (1988).

[15] 上坂：ニューラルネットと学習可能性，電子情報通信学会技術報告，*CAS 89-103, NLP 89-47* (1989), 69-74.

[16] 上坂，尾関：パターン認識と学習のアルゴリズム，文一総合出版，1990.

[17] 浦浜：ニューラルネットの最急降下学習法の収束速度，電子情報通信学会論文誌，**J 72-D-II** (1989), 298-301.

5月会合記録

5月14日 (火)	OR事例集編集委員会	10名
5月15日 (水)	庶務幹事会	9名
5月20日 (月)	OR誌編集委員会	14名
	FMESシンポジウム実行委員会	5名
5月21日 (火)	理事会	16名
5月24日 (金)	国際委員会	8名

第1回理事会議題

3—5—21

1. 平成2年度評議員会議事録の件
2. 平成2年度第7回理事会議事録の件
3. 平成2年度通常総会議事録の件
4. 入退会の件
5. 各支部総会報告の件
6. 平成2年度支部長会議開催報告・議事録の件
7. 平成2年度秋季研究発表会予算（案）の件
8. CEMIT-CECOIA 3協賛の件
9. 平成3年度委員会委員・幹事委嘱の件