ニューラルネットワークの学習過程の定量的理解

東北大通研　本堂 毅

概要

情報の自己組織的獲得過程のプロトタイプの1つとして、ニューラルネットワークモデルの学習過程を取り上げ、その学習過程に現れる普遍的な性質を議論する。特に、学習がある時間に急激に進む現象に着目し、この前後での系の時間発展を議論する。

1　序論

スピングラスに代表されるように、近年多安定系のダイナミックスが「複雑系」の数理の中心テーマとして脚光を浴びている。ニューラルネットワークのダイナミックスもこのような観点から活発な研究が行われている。情報処理に関して、多安定のダイナミックスが現れる典型例はニューラルネットワークであろう。いうまでもなく、Hopfield モデル[1]のダイナミックスは、スピン系との対応が容易であることから、統計力学的な解析が進んでおり[2]、最近では学習問題についても、同様の解析が行われている[3]。以下に見るように、ここでは、学習過程に興味を絞り、数値シミュレーションを通して得た結果を報告する[4]。数値シミュレーションは解析的な取扱が困難な問題でも取り扱うことが可能であり、解析的な取扱いのために、ともすると落としがちな本質に迫れるものと期待される。また、その時間発展を記述するのに適当なメジャーを見つけることは、後の理論的、解析的理解の基礎となるだろう。

2　モデルと背景

学習対象としては、低自由度でありながら単純ではない系列を生成する、という点から離散的カオス時系列を用いる (テントマップ)[5]。連続的時系列であっても低自由度であれば Farmer ら[6] が行っているように埋め込みの手法によって離散的時系列のマップの学習の問題に帰すことが出来る。学習アルゴリズムとしては、Rumelhart ら[7] によって提案され、現在最も応用例の多いバックプロパゲーションアルゴリズムを用いた。このアルゴリズムは、パラメーター (荷重) 空間に於けるトライアル・アンド・エラー法を最も数学的に簡潔な形で表現したもの、と考えられるからである。トライアル・アンド・エラー法は他の適応的アルゴリズムにも共通にみられる特徴と考えられる。

学習法は以下に示すように、1単位時間 (ステップ) 毎に1データが提示され、ネットワークはその時点での自乗誤差を減らすように、徐々に荷重を変えて行くものとする。これは確率的降下法[8] と呼ばれる時間に関してローカルな学習法である。ネットワークの構造は、4層のフィードファワード型を用いる。舟橋らによって、3層のネットワークを用いて任意の連続関数実現 (学習ではない) できることが証明されている[9]。テントマップは

$$x_{n+1} = f(x_n) = \begin{cases} rx_n & \text{for } x_n \leq 1/2 \\ r(1-x_n) & \text{for } x_n > 1/2 \end{cases} . \quad (1)$$

であり、x は入力、σ は出力である。ネットワークは

$$\begin{aligned} y_i &= tanh(\omega_i^1 x - \omega_0^{1i}), \\ z_i &= tanh(\sum_{j=1}^{3} \omega_{ij}^2 y_j - \omega_0^{2i}), \qquad (2) \\ \sigma &= \sum_{i=1}^{3} \omega_i^3 z_i, \end{aligned}$$

と書かれる。σ はネットワークの出力である。

自乗誤差は

$$E = (\sigma - f(x_n))^2/2, \qquad (3)$$

と書かれる。ここに、$f(x)$ はテントマップの関数関係: $x_{n+1} = f(x_n)$ である。学習アルゴリズムは

$$\vec{\omega}_{n+1} = \vec{\omega}_n + \delta\vec{\omega}_n, \qquad (4)$$

$$\delta\vec{\omega}_n = -\epsilon\frac{\partial E}{\partial\vec{\omega}}, \quad (\epsilon = 0.05). \qquad (5)$$

である。

3　結果と考察

まず、典型的な学習例を取り上げる。

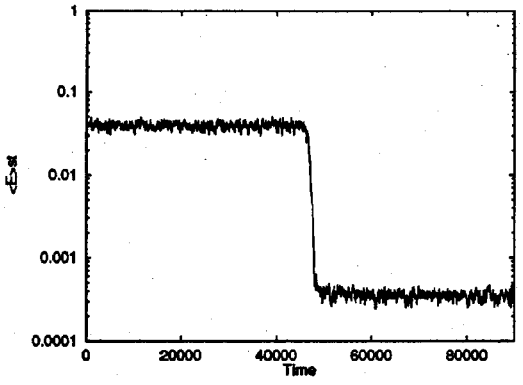

Fig.1, Temporal evolution of mean squared error, where $\langle\ \rangle_{st}$ is a short time average over 500 time step. t_{cr} is the critical time. Initial values of the weight vector, $\omega^n_{i,j}$ are independent uniform random numbers with range [-0.05,0.05], $r = 1.90$.

ここに見られるように、学習過程ではテントマップに限らず多くの場合、誤差が急激に減少する段階 (tcr) が存在する。これは数理心理学に於いてS字型学習曲線[10] として知られているものに対応する。従って、この段階の前後で系に質的な変化が起こっているものと考えらよう。ネットワークに獲得されているマップの時間発展を見ると、以下の様に tcr の前ではマップの平均値にフィッティングし、tcr で一気にマップの特徴を獲得していることが分かる。

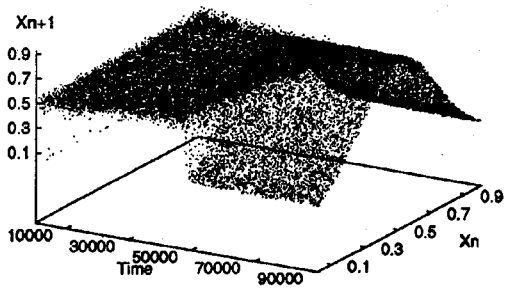

Fig.2, Temporal evolution of the map learned by the networks. Initial condition as in Fig.1.

学習の時間発展は荷重ベクトルの時間発展に等価である。そこで、荷重ベクトルの時間発展を３次元に落として見ると、一般に tcr 近傍で速度は著しく大きくなっていることがわかる。

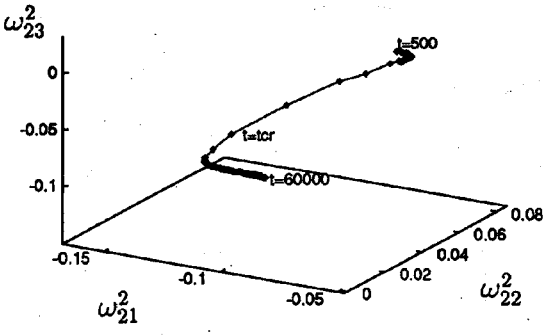

Fig.3, Time evolution of a weight vector for r=1.99. Data are plotted every 500 time steps. $\omega^n_{i,j}$ are independent random number of range [-0.05, 0.05], r=1.99.

この速度の増加の原因を探るため各ステップでの荷重ベクトルの変化量を調べると、tcr 近傍でその値は全く増加していないのである。

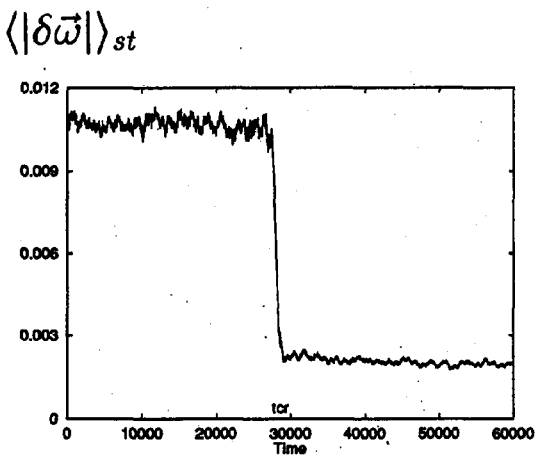

Fig.4, Time evolution of the velocity $\langle|\delta\vec{\omega}|\rangle_{st}$, calculated at each time step. The time-averaging smooths the line, but does not affect its shape. Initial condition as in Fig.3.

ミクロな動きがマクロな動きを作り出すもう1つの可能性は、系に何らかの(時間的な)秩序が生ずることだろう。そこで、荷重ベクトルのN次元空間における秩序を特徴づけるメジャーとしてコヒーレンス

$$C \equiv |\langle\delta\vec{\omega}\rangle_{st}|/\langle|\delta\vec{\omega}|\rangle_{st}, \qquad (6)$$

を定義した。$\langle\ \rangle_{st}$ は時間Tに関する短時間平均である(ここではT = 500)。コヒーレンスはTの値の選択にはあまり依存しない(Fig.5)。

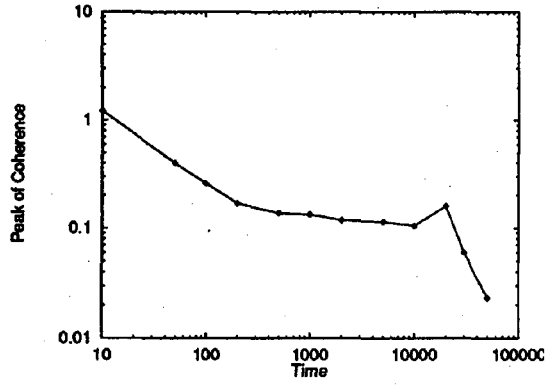

Fig.5, T-dependence of the peak value of the *coherence*. The *coherence* is not sensitive to the choice of T, the period of the short time average. Initial condition as in Fig.3.

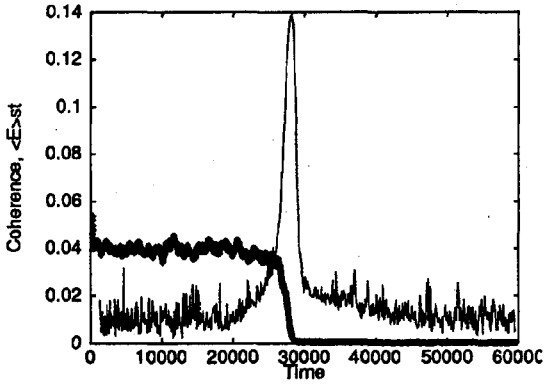

Fig.6, Time evolution of the *coherence* —— and the error —◇—. Initial condition as in Fig.3.

Fig.6は、コヒーレンスがtcrで鋭いピークをもつことを示している。これは、荷重空間の動きに時間的な秩序が形成されることを確かに示している。

ここでのコヒーレンスを用いた解析はややマクロな解析である。そこで、荷重ベクトルのステップ毎の変化 $\delta\vec{\omega}$ を特徴づけるために、以下のメジャー(ディレクションコサイン)を定義する。

$$\alpha \equiv \frac{\overrightarrow{orbit}\cdot\delta\vec{\omega}}{|\overrightarrow{orbit}||\delta\vec{\omega}|}, \qquad (7)$$

ここに、$\overrightarrow{orbit} \equiv \vec{\omega}_{t+T/2} - \vec{\omega}_{t-T/2}$ and T=200である。Fig.7, Fig.9に見られるように、$t \ll t_{cr}$ では α は1か−1近傍にのみ値を持つ。しかし、tcr近傍においてその中間領域に値を持つ確率が増えている。N次元空間における立体角を考えることにより、次元が上がれば上がる程、任意のベクトルに直交する空間が広がってくることが解かる。したがって、ここに見られる分布の変化は、荷重ベクトルの動く空間がtcr近傍で広がっていることの反映である、と解釈することが出来る。

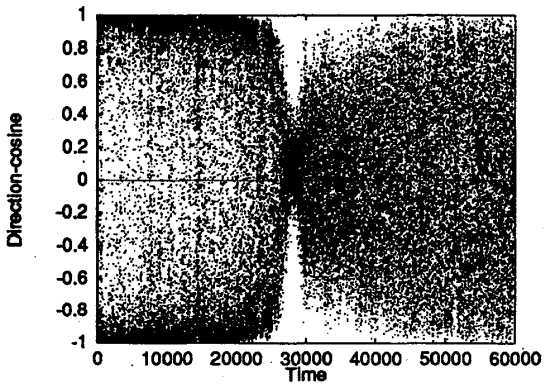

Fig.7, *Direction − cosine* of $\delta\vec{\omega}$ for each singal input. Initial condition as in Fig.3.

ここまでの結論は、入力が時系列ではなく、単に関数の入出力関係を学習するために、入力を例えばランダムに選んで提示している場合にも共通である。それでは、時系列特有の性質はどこに現れるのだろうか？実際、テントマップでr＝2としたときの時間相関関数$< x_i x_j >= (1/12)\delta_{ij}$ [5] は一様乱数[0,1]のそれに同じである(従って、パワースペクトルも等しい)。

しかしながら、テントマップ(r=1.999)から生成される決定論的時系列を入力として用いた場合と、一様乱数を用いた入力を選んだ場合では、２つの間には収束性に大きな違いが生じていた。その典型例をFig.8,
Fig.9に示す。

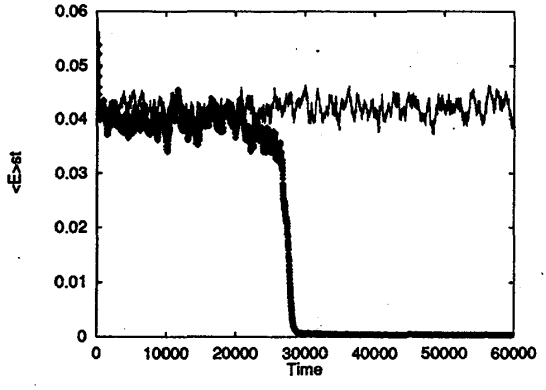

Fig.8, Typical difference of learning speed between uniform random numbers ——, and tent map chaos —◇—. Initial condition as in Fig.3.

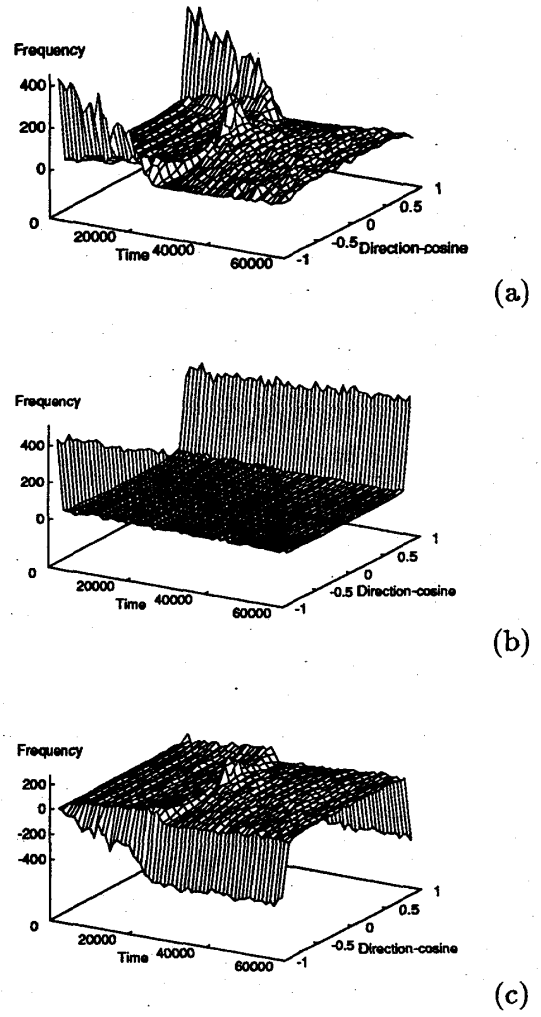

Fig.9, Frequency of *Direction − cosines* of $\delta\vec{\omega}$. The range of *Direction − cosine* [-1,1] was devided into 20 bins, and counted for each 1000 time steps. (a) Frequency for correlated inputs generated by a tent map. (b) Frequency for random number inputs. (c) Difference of frequencies between (a) and (b). (a) and (b) have the same initial condition as in Fig.3

ここで比較した双方の系においては、入力の普遍密度が等しい(一定)であるために、入力時系列無限大で定義される
$ES(\vec{\omega}) \equiv \langle E(\vec{\omega})\rangle_{T(\text{period of averaging})\to\infty}$ にも違いが無い。確率的降下法である、という点でも一様乱数もテントマップも同様である。従って、時系列の持つ(短時間)相関が学習に重要な役割を持つと結論された。無論、(５)式で$\epsilon \to 0$とすれば、短時間相関は消え、厳密な最急降

下法となるが、収束時間無限大をも意味する。これは、学習が環境との接触から達成されることの反映であり、Hopfield モデル[1)]等の想起過程とは本質的に異なる点を示している。

以上、学習過程をシミュレーションを通して眺めることにより、様々な特徴を見てきた。これらの現象は、テントマップに限らず多くの時系列あるいは、関数学習にも多く現れるようである。ここで定義したメジャーは、容易に他のモデルや(荷重ベクトルの)部分空間に適用出来るため、これを用いて、学習過程における普遍性の成り立つ範囲や由来を考えることにより、適応的学習の理解を進めて行きたいと考えている。

短時間相関や、学習曲線といったものは、旧来の統計的手法で扱うと、平均化操作の中で消えてしまう現象である。多安定系を研究する際共通する悩みとして、「初期値依存性が大きい」ということがあり、これが普遍的性質の理解の上で妨げになっているが、一方初期荷重に関するアンサンブルをとれば、鋭いカーブを持つ学習曲線もなだらかなものとなってしまう。ここに、多安定系の解析の難しさが現れている。「短時間相関」の学習における役割を考えることは、「複雑系」に於けるカオスの役割を考える上でも興味深い。TSP 問題をはじめとする様々な場面でニューラルネットワークにおけるカオスが研究され、興味深い結果が出ている[11-13)]が、その理由は十分明らかになっていないように思われる。普遍密度を同じにした時に残される「カオス」と「ノイズ(乱数)」の違いであるこの「短時間相関」は、マクロな統計量では消えてしまいがちな性質である。しかし、この違いが大きな役割を果たす系が存在することも明らかになった。今後、理論化を試みたい。

参考文献

[1] J.J.Hopfield: Proc. Natl. Acad. Sci. USA **81** (1984) 3088.

[2] D.J.Amit et al.: Phys. Rev. Lett. **55** (1985) 1530 等.

[3] H.S.Seung et al. : Phys. Rev. A **45** (1992) 6056.

[4] H.G.Schuster: *Deterministic chaos*, (Physik-Verlag, Germany, 1984).

[5] J.D.Farmer: Phys Rev Lett. **59** (1987) 845.

[6] D.E.Rumelhart et al.: *Parallel Distributed Processing*, (MIT press, 1986).

[7] K.Funahashi: Neural Networks. **2** (1989) 183.

[8] J.A.McGeoch et al.: *The psychology of human learning, 2nd ed.*, (Longmans, New York, 1952).

[9] S.Amari: IEEE Trans. Elect. Compu.-16 (1967) 299.

[10] H.Nozawa: CHAOS **2**, 3 (1992) 377.

[11] Y.Hayakawa et al.: 信学技報、NLP-37 (1992) 19.

[12] M.Inoue et al: Phys. Lett. **A 158** (1991) 373