

ニューラルネットワークによる力学系の学習過程の非線形解析

1995年12月

中 島 弘 之

概要

階層型ニューラルネットワークによる力学系の学習過程は教師信号に駆動されたパラメータ空間上の状態変化である。これは、典型的な高次元非線形ダイナミクスであるため、学習過程の理論的解析は極めて困難となる。しかし、力学系の学習を基本とする時系列予測システムなどの基礎理論として、そのダイナミクスの解明は重要であり、また、非線形力学系の基礎理論の観点からも非常に興味深い問題である。

本研究は、階層型ニューラルネットワークによる力学系の学習に関して、その動的過程と収束性を力学系理論に基づき数理解析的に考察したものである。

はじめに、ニューラルネットワークによる離散力学系および連続力学系の学習過程を確率的降下法の理論を用いて統計的に解析した。まず、学習のアルゴリズムを、時間波形そのものの学習と、ベクトル場あるいは変換を表わす関数の学習に分類し、関数の学習は確率的には逐次型のバックプロパゲーション法と等価であること、時間波形の学習も近似的にはバックプロパゲーション法と一致することを証明した。また、この理論結果を数値実験によって検証し、学習初期の速度について理論が妥当であることを確認した。さらに数値実験において、カオス時系列を学習する場合に学習が急激に加速する現象を見出し、これについて考察を加え、アトラクタの対称性および学習されたカオスとバックプロパゲーション法の持つ統計的性質が加速の原因となっていることを解明した。

次に、力学系学習の最適値への局所収束性を、結合荷重パラメータを変数とする非自律的力学系の不動点の安定性として解析した。まず、教師信号が周期的、概周期的、カオス的な場合のそれぞれについて、学習の局所収束条件を導出し、種々の定理のかたちで証明した。また、学習の最適値近傍での収束過程の数値計算を行ない、収束速度と固有値、リアプノフ指数との関係を考察した。最後に、バックプロパゲーション法を含む一般の逐次修正型の勾配降下法により二乗誤差を最小化するアルゴリズムによるカオスの学習では、最適状態の引力圏が非常に複雑になる場合があることを指摘し、そのような引力圏の存在証明を行ない数値実験例を示した。

本研究で得られた知見はニューラルネットワーク技術の実用面において重要な理論的基盤となるだけでなく、カオス力学系の研究においても「力学系としての学習」という新たな問題を提起するものである。

目次

第1章 序論	1
1.1 まえがき	1
1.2 ニューラルネットワーク研究の概説と本研究の位置付け	4
第2章 リカレントニューラルネットワークの構造と力学系の学習則	7
2.1 緒言	7
2.2 リカレントニューラルネットワークの構造とダイナミクス	7
2.3 離散力学系の学習則	9
2.4 連続力学系の学習則	12
2.5 関数近似としての力学系学習	14
2.6 連続力学系の学習則の分類	16
2.7 結言	18
第3章 力学系の学習過程の確率的降下法の理論に基づく考察	19
3.1 緒言	19
3.2 確率的降下法としての学習則の定式化	19
3.3 離散力学系の学習	21
3.3.1 BP法による学習	21
3.3.2 関数誤差法による学習	21
3.3.3 軌道誤差法による学習	22
3.4 連続力学系の学習	22
3.4.1 速度誤差法による学習	23
3.4.2 波形誤差法による学習	24
3.5 適応制御型学習則の収束性	25
3.6 結言	27
第4章 力学系の学習実験	28
4.1 緒言	28
4.2 離散力学系の学習	28

4.2.1	学習曲線と学習に要する時間の定義	29
4.2.2	関数誤差法による学習過程	30
4.2.3	軌道誤差法による学習過程	31
4.2.4	軌道誤差法における学習の加速	36
4.2.5	学習の加速の発生メカニズム	41
4.2.6	結合荷重の収束過程	52
4.3	教師信号の提示方法について	59
4.3.1	有限時間時系列データを用いた学習	60
4.3.2	時系列の分布密度・時間相関を変化させた学習	67
4.4	連続力学系の学習実験	69
4.4.1	van der Pol 方程式の学習	69
4.4.2	Rössler 方程式の学習	77
4.5	結言	78
第 5 章	差分方程式の理論を用いた学習収束過程の解析	83
5.1	結言	83
5.2	離散力学系の学習の収束性	84
5.2.1	学習を表わす差分方程式	84
5.2.2	二乗誤差を最小化する勾配降下法の収束条件	85
5.3	連続力学系の学習の収束性	100
5.3.1	学習を表わす差分方程式	100
5.3.2	最適制御型学習則の収束条件	101
5.4	結言	111
第 6 章	学習収束過程と最適状態の引力圏	114
6.1	結言	114
6.2	数値実験	114
6.2.1	離散力学系学習の収束過程	114
6.2.2	法線リアブノフ指数と最適点の安定性	118
6.2.3	最適点が不安定化したときの学習の最終状態	124
6.3	勾配降下法による力学系の学習に現れる riddled basin	127
6.3.1	riddled basin を持つ学習系の構成と存在証明	127
6.3.2	学習係数をカオス的に動振した勾配降下法	132

6.4 結言	132
第7章 結論	135
7.1 本研究の成果	135
7.2 今後の課題	137
謝辞	138
付録A リカレントニューラルネットワークの学習則の導出	139
A.1 離散力学系の学習則	139
A.1.1 関数誤差法	140
A.1.2 軌道誤差法	141
A.2 連続力学系の学習則	142
A.2.1 速度誤差法	143
A.2.2 波形誤差法	144
付録B リアブノフ指数と法線リアブノフ指数	145
付録C 法線リアブノフ指数と引力圏の測度に関する考察	147
参考文献	152

第1章 序論

1.1 まえがき

自然科学の本質は現象の予測である。ニュートン力学では、閉じた系のすべての状態変数の初期条件がわかれば、その後の系の状態は運動方程式によって決定される。すなわち、十分な情報があれば現象の予測は理論的には可能であるとされていた。この認識は、量子論の発見によって、系の状態を任意の精度で観測することが不可能であることが示されても本質的には変化を受けなかった¹⁾。それは、自然科学の考察の対象となる定常状態は軌道安定であり、定常状態を観測するがぎり、わずかな測定誤差は時間とともに減少することはあっても拡大することはないという経験に基づく信念があったためと思われる。つまり、誤差が不可避であっても、その大きさに粗視化した系を考えれば十分意味のある予測が原理的にはできるものと信じられていたのである。

しかし、近年、定常的であるが軌道不安定な現象が普遍的に存在することが明らかにされた。このような現象では、初期条件の誤差をどれだけ小さくしても、その誤差は指数関数的に増加するため、瞬く間に観測にかかるオーダーに拡大する。すなわち、長期的な予測が原理的に不可能になる。このような現象は、数学的に言えば、軌道の全体（不変集合）は漸近安定であるが、個々の軌道は反発しあい、指数関数的に離れていくような状態変化を奏しており、ゆるやかな定義で「カオス」と呼ばれている^{2)・3)}。指数関数的に離散していく軌道が再び接近し、全体として有界な不変集合を構成する原因は現象を記述するシステムの非線形性にある。この非線形性のため一般にカオスはランダムノイズとほとんど区別のつかない非常に複雑な振舞いを見せる。

このようにカオスの発見により、自然現象の長期予測の困難さについて、その本質が数学的に説明可能となった。しかし、逆に考えれば、自然現象や社会現象にみられる複雑な時系列の多くがカオスであり、不確定な要因を含まない決定論的な法則によって生み出されているならば、従来ランダムとみなされ統計的な処理しかできなかった時系列を決定論的に扱うことが可能になる。そこで、これまで統計的な手法で行われてきた時系列予測に、力学系理論^{2)・15)}に基づく決定論的手法を適用しようという考えが生まれてきた^{16)・19)}。カオス的な現象を長期的に予測する

ことは本質的に不可能であるが、それを生み出す規則を抽出すれば短期的な予測は可能であるという性質を利用しようというものである。いかに複雑な時系列であっても、或る有限次元の力学系によって生成されるものであれば、その力学系を同定することにより任意の初期状態を起源とする現象を予測することが可能になる。有限次元の力学系は、連続系ならば常微分方程式、離散系ならば差分方程式で記述される。したがって、力学系の同定はこれらの方程式に現れる関数を近似することに帰着する。関数近似の手法は種々提案されている。ニューラルネットワークを用いる手法はその代表的なものの一つである^{19)・24)}。

ニューラルネットワーク (Neural Network: 以下、NNと略す) は生物の神経回路を模倣した抽象的回路網であり、ユニットと呼ばれる多数の素子が加重結合され信号をやり取りすることにより、連想記憶や最適化問題の求解などの情報処理を行なうことができる^{35)・41)}。中でも階層型のNNはノンパラメトリックな関数近似系として、次のような特長を持つ。まず、ユニット数が十分にあれば原理的には任意の連続関数を近似できることが証明されている²⁰⁾。また、バックプロパゲーション法 (Back-Propagation: 以下、BP法と略す)⁴²⁾ およびそれを拡張した学習アルゴリズムにより、時系列データに基づきネットワークの結合荷重を順次修正して近似関数を構成することが可能である。このように、階層型NNは優れた関数近似能力とBP法という有効な学習方法を持つことから、力学系の同定のための関数近似系として有望視されている。

このような期待のもとに、NNを時系列の学習および予測に利用する研究が現在活発に行われており、カオスの工学的応用とNN研究の両分野において重要なテーマの一つになっている。

本論文は、以上のような階層型NNによる力学系の学習の研究の流れにおいて、

- (1) 力学系、特にカオス力学系の学習過程を現象論的に考察する、
- (2) 結合荷重空間における状態変化を表わす力学系として学習を捉える、

という二つの観点から力学系学習の数理的性質に関する知見を与えることを目的として述行した著者の研究^{101)・134)}をまとめたものである。本研究では、力学系の学習過程に生じる現象を計算機実験により観察し、従来の数理的理論の適用可能性を吟味すること、さらに、カオス力学系の解析手法を適用することにより、学習過程と学習の収束に関するより厳密な理論を構築することを目的とした。力学系

ここで述べるような情報処理の様式としてのニューラルネットワークを生体の神経回路網と区別するため、人工ニューラルネットワーク (Artificial Neural Network) と呼ぶこともある。

の学習に関する研究は数多くなされてきたが、非線形力学系の理論を用い統計的操作を極力排除することにより、その学習収束性を厳密に解析した例は筆者の知るかぎり本研究において他にない。

本論文の構成を以下にまとめておく。

次節1.2では、階層型NNの研究の歴史を概観し、その流れにおける本研究の位置付けを明確にする。

第2章では、本研究で主に対象とする3層階層型リカレントニューラルネットワーク (Recurrent Neural Network: 以下、RNNと略す) の構造とダイナミクス、その学習則を説明する。まず、離散力学系の学習則として関数誤差法と軌道誤差法を、次に連続力学系の学習則として速度誤差法と波形誤差法を導き、それらの特徴を述べる。また、本研究では扱わない他の学習法も簡単に紹介し、本研究で用いる手法との関連について述べる。

第3章では、甘利によって構築された確率的降下法の理論^{43, 44)}を第2章で述べたRNNの学習アルゴリズムの解析に適用し、BP法との関係を明らかにし、学習速度の近似的評価式を導出する。また、確率的降下法に対する「学習収束の定理^{43, 44)}」を連続力学系の適応制御型学習に拡張する。

第4章では、離散および連続力学系の学習の計算機実験結果について述べる。まず、1次元離散力学系の学習実験を行ない、周期軌道および不安定性の小さいカオスの学習においては、BP法と関数誤差法が平均的には等価であることを示し、確率的降下法の理論に基づく学習速度の評価が妥当であることを示す。また、前記の学習実験において観測されたカオス学習の加速現象について述べ、その原因について詳細な考察を行なう。次に、教師信号の提示方法が学習に及ぼす影響を、離散力学系の学習を対象として考察する。最後に、連続力学系の学習実験結果について述べ、離散系の場合と同様に確率的降下法による解析の正当性を検証する。

第5章では、RNNによる学習を含む、より一般の勾配降下法に基づく力学系の学習を、学習系の結合荷重空間における状態変化、すなわちそれ自身一つの力学系であると考え、学習の収束条件をこの力学系の不動点の安定性として解析する。解析には、不動点の近傍での線形近似系の零点の安定性から元の不動点の安定性を推定するリアブノフの第一の方法⁴⁵⁾を使用する。この手法により、教師信号である時系列が、周期的、概周期的、カオス的な各場合について、それぞれの学習の収束条件を明らかにし、種々の定理のかたちで証明する。

第6章では、学習の最適点の近傍での収束過程の数値計算を行ない、最適点の固有値、リアプノフ指数と収束速度との関係を考察する。学習は教師信号に駆動された非自律的力学系である。このような非自律系、特に、カオス的な駆動項をもった系の不動点の安定性については、従来ほとんど研究がなされていなかったが、最近、カオスの同期^{46)・47)}や riddled basin^{48)・52)} といった新しい現象が発見され、その発生条件との関連から重要視されている問題である。本研究は、NNによる学習を結合荷重空間における教師信号に駆動された力学系と考える観点からのこの問題への一つのアプローチでもある。本章では、NNによる学習における riddled basin の発生についても考察を行ない、最適状態が riddled basin をもつような逐次的勾配降下法による学習系のモデルの構成を試みる。

第7章では、本研究で得られた成果をまとめ、今後の課題を述べる。

1.2 ニューラルネットワーク研究の概説と本研究の位置付け

ニューラルネットワーク (NN) は生体の神経回路網を模倣した情報処理系である。NNの構成要素はユニットと呼ばれる抽象的素子であり、これは、生体の神経細胞、ニューロンに相当する。各ユニットは複数の入力信号を受信し、これに変換を施して一つの信号を出力する機能を持つ。個々のユニットの信号変換機能は比較的単純であるが、このようなユニットが多数結合し、信号を相互に伝達することにより、NNは全体として複雑なダイナミクスを有している。

NNはその構造により、階層型(層状)ネットワークと相互結合型ネットワークに大別される。階層型ネットワークは、ユニット全体が複数の層状のグループに分かれ、隣接した層に含まれるユニット同士のみが結合したネットワークである。信号は一つの層のユニットに入力され、隣接した層の間を伝搬し、末端の層から出力される。この伝達は一方方向のみに行なわれ、信号の入力される層を「入力層」、出力される層を「出力層」と呼ぶ。その他の層は「中間層」あるいは「隠れ層」と呼ばれている。一方、相互結合型のネットワークは、一般にすべてのユニットが相互に結合し、互いに信号をやり取りする構造を持っている。階層型ネットワークは、パターン認識、関数近似などに利用され、相互結合型ネットワークは、主に連想記憶、最適化問題の求解などに利用されている。

本論文では階層型ネットワークを研究の対象とするため、以下、その研究の歴史と現状を概観する。

階層型ネットワークの研究は1958年に発表されたRosenblattのパーセプトロンに

端を発する³³⁾。パーセプトロンは3層構造を持ったネットワークで、提示パターンを入力層から信号として取り込み、これにしきい値関数と加重線形和に基づく非線形変換を施し、0と1のパターンとして出力する変換装置である。中間層から出力層への結合の荷重を調整することにより、入力信号を適当なパターンに分類できるような学習方法が与えられている。また、線形分離可能なパターンの識別学習は、有限回で収束することが証明された(パーセプトロンの収束定理)³⁴⁾。しかし、パーセプトロンは中間層から出力層への結合荷重のみが調節可能であり、このような単純な構造・学習則の限界がMinskyとPapertによって指摘された。彼らはパーセプトロンのような単純な構造の学習機械で高度な知的情報処理を行なうことは不可能であると結論づけたのである³⁵⁾。

MinskyとPapertの指摘により、NNの研究の流行はしばらく下火になっていた。しかし、1980年台に入り、多層階層型ネットワークにおいて、すべての結合荷重が調節可能な学習則であるバックプロパゲーション法(BP法)がRumelhartらによって提案されたことにより、NNは再び注目を集めることとなった³⁶⁾。

BP法は出力信号の誤差を勾配降下法に基づき最小化するアルゴリズムであり、勾配の計算に微分の鎖法則を利用するものである。それは、最適化問題の解法のテクニックとして新しいものではなかった⁴⁰⁾。しかし、そのアルゴリズムの明快さと、記号処理的手法に基づく人工知能研究の限界が指摘されていたという時代背景、電子計算機の性能の急激な向上によりNNの大規模シミュレーションが可能となったことなどが相俟て、NN研究再燃の火付け役となったのである。BP法は階層型ネットワークによる学習の基本アルゴリズムとして画像・音声などのパターン認識問題に幅広く応用された。現在では階層型ネットワークをBPネットワーク、あるいは多層パーセプトロンと呼ぶことがある。

パーセプトロンの原型は出力が01パターンという離散値であったが、これを連続値とすることにより、階層型ネットワークを「連続関数のシミュレータ³⁷⁾」とみなすことができる。その関数近似能力については、任意の n 次元ユークリッド空間のコンパクト集合から m 次元空間への連続写像が任意の精度で近似可能であることが船舶によって証明された²⁰⁾。また、BP法の本質は勾配降下法であるから、階層型だけでなく入出力信号を有する相互結合型のネットワークの学習にも拡張できる。個々のユニットがダイナミクスを持つような相互結合型ネットワークはリカレントニューラルネットワーク(RNN)と呼ばれ、時間波形や時系列パターンの認識・獲得に利用可能である。BP法を拡張し、RNNによって時系列の学習を行

なうアルゴリズムは、佐藤²³⁾、Pearlmutter²⁶⁾、Williams と Zipser²⁷⁾らによって提案され、時系列予測やシステム同定などに応用されている。一般に RNN はすべてのユニット間の結合を許可するものであるが、特に階層型ネットワークの出力層から入力層へフィードバック結合を施した構造の RNN は、前述のように入出力関係が任意の連続関数を近似できることから、時系列を生成する力学系の同定に都合が良い。この構造の RNN によるシステムの同定・時系列の予測は、Lapedes と Farber¹⁹⁾、安達^{21, 22)}、松葉³⁰⁾らによって研究されている。

以上、概観したように、NN による力学系の学習に関する研究は、階層型 NN の学習アルゴリズム (BP 法) が開発され、また、その関数近似能力が理論的に保証されたことがきっかけとなり、フィードバック付き階層型 NN を用いる手法を中心として、ここ数年の間に急速に発展した。現在では、種々の産業分野への応用・実用化に向けた研究が活発に行なわれている。

さて、このように NN の研究はここ 10 年の間に急激な発展を遂げたのだが、1940 年代に始まり、Minsky と Papert の指摘の後も地道に続けられた膨大な数理解析的研究がその基盤にあることを忘れてはならない。特に、甘利は数理統計的手法を駆使し、NN 研究全般にわたる数理解析理論を構築した³⁵⁾。階層型 NN に関するものでは、NN を含むより一般の学習識別系を扱う確率的降下法の理論がある^{43, 44)}。この理論は、逐次型の BP 法による学習、すなわち、個々の提示パターンに対する誤差を評価し、結合荷重の更新を逐次的に行なう学習が最適状態に確率的な意味で収束することを保証する。また、学習の平均速度、最適点近傍でのパラメータの共分散行列の算出も可能にする。NN 研究の爆発的な流行が落ち着きを見せ始めた現在、実用化を急ぐあまりにややなおざりにされた感のあるこのような基礎研究の重要性が再認識されつつある。

以上のような研究の流れを踏まえ、本研究は、時系列、特にカオス時系列の NN による学習過程に着目し、カオス力学系の観点から、従来の理論の適用可能性を吟味し、さらに新たな理論の構築を目的として遂行した。特に NN の学習をカオス力学系の観点から解析する手法は、力学系理論を NN の研究に応用するに際しての一つの指針を与えるものと考えられる。

第2章 リカレントニューラルネットワークの構造と力学系の学習則

2.1 緒言

本章では、本論文で扱う三層階層型リカレントニューラルネットワークの構造とダイナミクス、およびこれを用いた力学系の学習アルゴリズムを連続系、離散系の場合に分けて説明する。

まず、三層階層型RNNの構造とダイナミクスを述べ、本論文で主に扱う学習則である、離散系を対象とした関数誤差法と軌道誤差法、連続系を対象とした速度誤差法と波形誤差法について、その原理、具体的な計算方法を説明する。次に、力学系の学習を関数近似と捉える観点からこれらの方法の相違点を整理し、BP法との関連を述べる。また、本論文では対象としないものも含めて、連続力学系の学習則を簡単に分類・整理する。

2.2 リカレントニューラルネットワークの構造とダイナミクス

本論文では、図2-1に示すような構造のRNNを対象とする²³⁾。このネットワークは、三層階層型NNの入出力層を同一視した構造をもち、入出力層に相当する n 個のダイナミックユニット（以下 DU と略記する）と、中間層に相当する M 個のシグモイドユニット（以下 SU と略す）とから構成される。 DU の出力信号は、離散系の場合は(2.1)の差分方程式、連続系の場合は(2.2)の微分方程式にそれぞれしたがって、RNNのダイナミクスを規定する。

$$\mathbf{x}(k+1) = \hat{f}(\mathbf{x}(k), \mathbf{w}) \quad (\mathbf{x} \in \mathbb{R}^n), \quad (2.1)$$

$$\frac{d\mathbf{x}(t)}{dt} = \hat{f}(\mathbf{x}(t), \mathbf{w}) \quad (\mathbf{x} \in \mathbb{R}^n). \quad (2.2)$$

ここで、 \mathbf{x} は DU の出力を表わす n 次元ベクトルであり、 k は離散時間($k \in \mathbb{Z}$)、 t は連続時間($t \in \mathbb{R}$)を表わす。 \mathbf{w} は次に述べる結合荷重を表わすパラメータを成分にもつベクトルである。関数 $\hat{f}(\mathbf{x}, \mathbf{w})$ も n 次元ベクトル関数であり、成分別に書くと次のようになる。

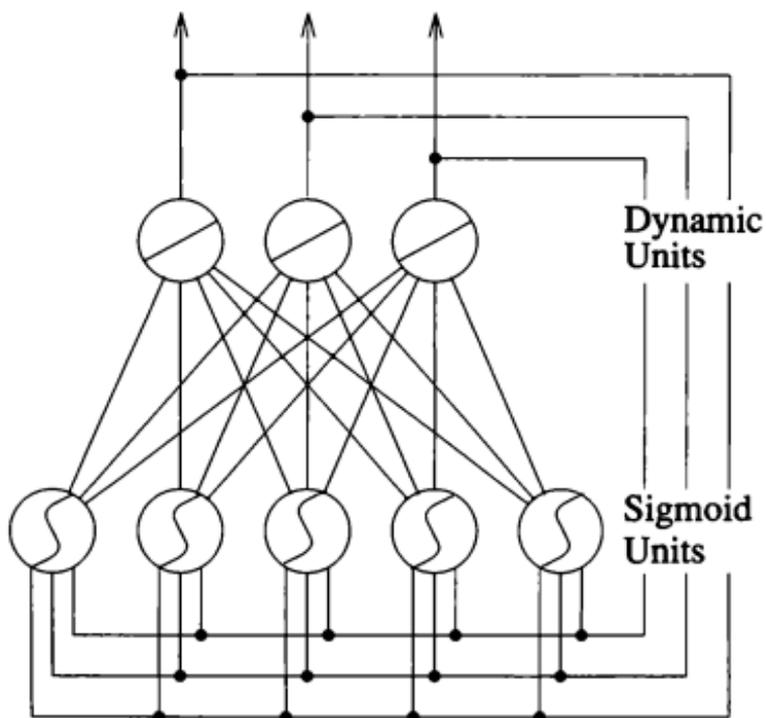


図 2-1 階層型リカレントニューラルネットワークの構造

$$\begin{aligned} \hat{f}_i(\mathbf{x}, \mathbf{w}) &= \sum_{m=1}^M W_{im}^A G(y_m) \quad (i = 1, \dots, n), \\ y_m &= \sum_{j=1}^n W_{mj}^B x_j + W_m^D \quad (m = 1, \dots, M). \end{aligned} \quad (2.3)$$

ここで、 y_m は m 番目の SL への入力を表わし、 $G(y_m)$ はその出力である。 W_{im}^A は m 番目の SL から i 番目の DL への、 W_{mj}^B は j 番目の DL から m 番目の SL への結合の重みを表わすパラメータであり、 W_m^D は m 番目の SL のバイアス項である。式(2.1)と(2.2)におけるベクトル \mathbf{w} は、

$$\mathbf{w} = \{W_{11}^A, \dots, W_{nM}^A, W_{11}^B, \dots, W_{Mn}^B, W_1^D, \dots, W_M^D\}$$

である。 $G(x)$ はシグモイド関数であり、本論文では $G(x) = \tanh x$ とする。

式(2.3)はまとめて書くと、

$$\hat{f}_i(\mathbf{x}, \mathbf{w}) = \sum_{m=1}^M W_{im}^A G\left(\sum_{j=1}^n W_{mj}^B x_j + W_m^D\right) \quad (i = 1, \dots, n) \quad (2.4)$$

となる。この関数は、 R^n のコンパクト集合から R^n への任意の連続関数を任意の精度で近似できることが証明されている²⁰⁾。

このネットワークとその連続系学習則は村上・佐藤²³⁾によって提案され、ローレンツアトラクタなどの複雑な時系列が学習できることが示された。時系列の学習をその時系列を生成する力学系の関数の学習と考える立場からみれば入出力関係の見通しが良く、また任意の連続関数を近似できることが保証されているので、力学系の学習に適した RNN であると考えられる。

2.3 離散力学系の学習則

学習対象として次式のような n 次元離散力学系を考える。

$$\mathbf{q}(k+1) = f(\mathbf{q}(k)) \quad (\mathbf{q} \in R^n). \quad (2.5)$$

以下、離散系の学習則を説明する。これは、連続系の学習を対象とした、村上・佐藤²³⁾の学習則を離散時間に書き直したものである。

・シグモイド関数は、一般には、有界かつ単調非減少な C^1 級の関数と定義されている²⁰⁾。

[Step 1] $x(0) = q(0)$ とし、時刻 $k=0$ から $N_B - 1$ まで RNN を式 (2.1) に従って動作させ、誤差評価関数 E を算出する。

[Step 2] 勾配降下法に基づき、 E を減少させるようにパラメータ W^A 、 W^B 、 W^D を更新する。

[Step 3] 時刻 $k = N_B$ を改めて時刻 0 とみなし、[Step 1] に戻る。

この3つのステップを1サイクルとし、誤差が小さくなるまでこのサイクルを繰り返すことにより学習を進める。

[Step 1] において、 $x(k)$ の値を $q(k)$ の値に拘束する方法と拘束しない方法の二通りの方法が考えられる。本論文では、前者を「関数誤差法」、後者を「軌道誤差法」と呼ぶことにする。両者は誤差評価関数の定義が異なり、それぞれ次式の E_1 、 E_2 のようになる。

$$E_1(w) = \frac{1}{2} \sum_{k=0}^{N_B-1} \|\hat{f}(q(k), w) - f(q(k))\|^2, \quad (2.6)$$

$$E_2(w) = \frac{1}{2} \sum_{k=0}^{N_B-1} \|\hat{f}^{k+1}(q(0), w) - f^{k+1}(q(0))\|^2. \quad (2.7)$$

ここで、 $\|\cdot\|$ はユークリッドノルムを表わす。

E_1 と E_2 の違いは図 2-2 のように表現される。すなわち、関数誤差法では RNN のダイナミクス (2.1) は E_1 の算出のためにだけ用い、 $x(k)$ の値は $q(k)$ の値に拘束され続ける。これは、教師信号 q の取りうる値に定義域を制限された関数 f を学習していることを意味する。一方、軌道誤差法では、 $x(0) = q(0)$ の初期条件のもとで式 (2.1) にしたがって RNN を自由に動作させた場合の $k=0$ から $k = N_B - 1$ までの軌道の二乗誤差の累積を評価関数 E_2 としている。

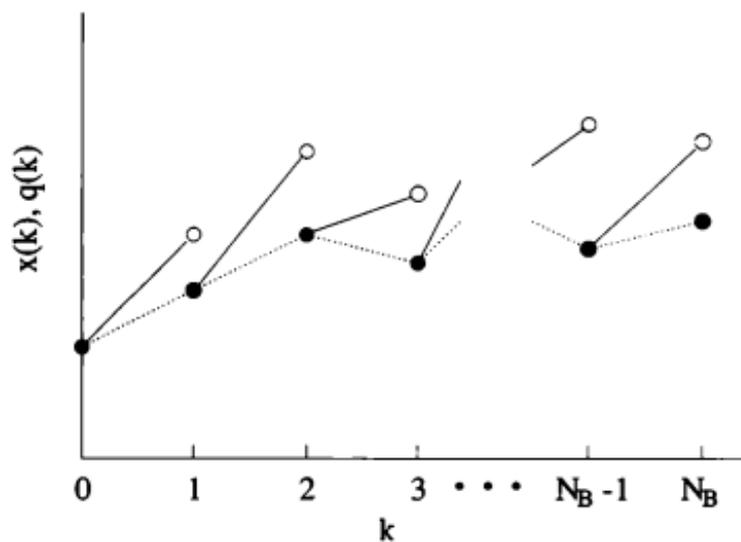
結合荷重 w の更新は、二つの方法とも次の形の勾配降下法にもとづいて実行する。

$$w \leftarrow w - \eta \nabla_w E_l \quad (l=1,2). \quad (2.8)$$

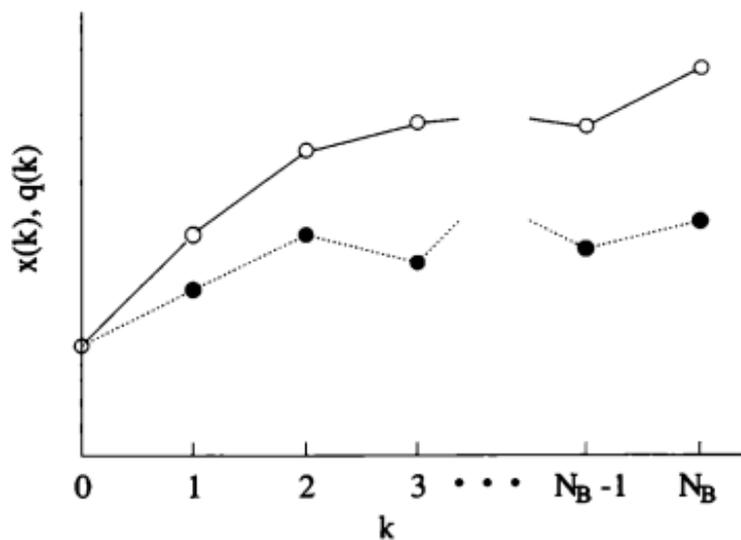
ここで、 ∇_w は w に関する勾配 (gradient) を表わす。また、 η は学習係数と呼ばれる正数である。 $N_B = 1$ のときは、ともに関数 f を BP 法、

$$w \leftarrow w - \frac{1}{2} \eta \nabla_w \|\hat{f}(q(0), w) - f(q(0))\|^2$$

で学習していることになる。



(a) 関数誤差法



(b) 軌道誤差法

図 2-2 関数誤差法と軌道誤差法 (○: RNNの出力 $x(k)$ 、●: 教師信号 $q(k)$)

パラメータ w の修正量を $\Delta w = -\eta \nabla_w E_i$ と表わすと、 w の各成分の修正量は次のように計算できる。

$$\Delta W_{im}^A = -\eta \sum_{k=0}^{N_g-1} p_i^X(k) G(y_m(k)). \quad (2.9)$$

$$\Delta W_{mi}^B = -\eta \sum_{k=0}^{N_g-1} p_m^Z(k) G'(y_m(k)) x_i(k). \quad (2.10)$$

$$\Delta W_m^D = -\eta \sum_{k=0}^{N_g-1} p_m^Z(k) G'(y_m(k)) \quad (i = 1, \dots, n, \quad m = 1, \dots, M). \quad (2.11)$$

ここで、 p^X と p^Z はラグランジュの未定乗数であり、以下のように計算される。まず、 p^Z は関数誤差法、軌道誤差法ともに次式で与えられる。

$$p_m^Z(k) = \sum_{j=1}^n p_j^X(k) W_{jm}^A \quad (m = 1, \dots, M). \quad (2.12)$$

p^X は、関数誤差法の場合は次式で与えられ、

$$p_i^X(k) = \sum_{m=1}^M W_{im}^A G(y_m(k)) - q_i(k+1) \quad (i = 1, \dots, n). \quad (2.13)$$

軌道誤差法の場合は、次の逆時間方向の差分方程式にしたがって計算される。

$$p_i^X(k-1) = \sum_{m=1}^M p_m^Z(k) G'(y_m(k)) W_{mi}^B + x_i(k) - q_i(k),$$

$$p_i^X(N_g) = 0 \quad (i = 1, \dots, n). \quad (2.14)$$

以上の学習則の導出過程は付録 A に記す。

2.4 連続力学系の学習則

本節では、学習対象として次式のような n 次元連続力学系を考える。

$$\frac{dq(t)}{dt} = f(q(t)) \quad (q \in \mathbb{R}^n). \quad (2.15)$$

** 正確には、逐次型 B P 法である。一般のパターン識別問題では、B P 法は、教師信号である学習パターンを一つ提示するごとに結合荷重を更新する逐次型と、全パターンを提示してから更新する一括型の二つありの方法に分けられる。本論文では、特に断らない限り B P 法といえは逐次型のものを指す。関数誤差法と一括型 B P 法との関係については、2.5 節を参照されたい。

ここで f は R^n 上の C^1 -級関数とする。この系を式(2.2)の連続時間ダイナミクスを持った RNN で学習する。

以下の論述では、(2.15)の解で初期値 ($t=0$ での値) が q_0 であるものを $q(q_0, t)$ と表わす。同様に、 x_0 を初期値とする (2.2) の解を $x(x_0, t, w)$ で表わす。また、誤解のない場合は、 $q(t)$ を $q(q_0, t)$ の、 $x(t)$ を $x(x_0, t, w)$ の略記として用いることがある。

学習の基本アルゴリズムは以下のように前述の離散力学系の場合とはほぼ同様であり、1学習サイクルを連続時間で $t=0$ から T_B までとする点のみが異なる。

[Step 1] $x(0) = q(0)$ とし、時刻 $t=0$ から T_B まで RNN を (2.2) 式に従って動作させ、誤差評価関数 E を算出する。

[Step 2] 勾配降下法に基づき、 E を減少させるようにパラメータ W^A 、 W^B 、 W^D を更新する。

[Step 3] 時刻 $t = T_B$ を改めて時刻 0 とみなし、[Step 1] に戻る。

連続系の学習の場合も、誤差評価関数の違いによって二通りの学習則が導ける。ここでは、両者を速度誤差法と波形誤差法と呼ぶことにする。

速度誤差法は、RNN の入力の値を教師信号 $q(t)$ の値に拘束しながら、RNN の出力と教師信号の速度の二乗誤差を時間 $t=0$ から $t=T_B$ まで積分したものを誤差評価関数 E_3 とする。すなわち、

$$E_3(w) = \frac{1}{2} \int_0^{T_B} \|\dot{f}(q(t), w) - f(q(t))\|^2 dt \quad (2.16)$$

である。このように、速度誤差法で学習する場合には、RNN は微分方程式(2.2)に従う本来のダイナミクスを持たず、入力値である教師信号 $q(t)$ に対して出力 $\dot{f}(q(t), w)$ を算出する変換器として動作する。学習はこの変換を表わす関数 \dot{f} で関数 f を近似していくことになる。

一方、波形誤差法では RNN の初期値は教師信号のそれに一致させるが、その後 RNN を自己のダイナミクスで動作させ、一定時間間隔における出力と教師信号との誤差を評価する。すなわち次式の E_4 を誤差評価関数とする。

$$E_4(w) = \frac{1}{2} \int_0^{T_B} \|x(q_0, t, w) - q(q_0, t)\|^2 dt. \quad (2.17)$$

1学習サイクルでのパラメータの修正量は、両方法とも次式で計算される。

$$\Delta W_{im}^A = -\eta \int_0^{T_B} p_i^X(t) G(y_m(t)) dt, \quad (2.18)$$

$$\Delta W_m^B = -\eta \int_0^{T_B} p_m^Z(t) G'(y_m(t)) x_i(t) dt. \quad (2.19)$$

$$\Delta W_m^D = -\eta \int_0^{T_B} p_m^Z(t) G'(y_m(t)) dt \quad (2.20)$$

$$(i = 1, \dots, n, \quad m = 1, \dots, M).$$

ここで、 p^X と p^Z はラグランジュの未定乗数であり、速度誤差法では、 p_i^X は次のように定義される。

$$p_i^X(t) = \sum_{m=1}^M W_{im}^A G(y_m(t)) - \frac{dq_i(q_0, t)}{dt} \quad (i = 1, \dots, n). \quad (2.21)$$

波形誤差法では、 p_i^X は終端条件をもった次の微分方程式の解として定義される。

$$\frac{dp_i^X(t)}{dt} = - \sum_{m=1}^M p_m^Z(t) G'(y_m(t)) W_{im}^B - x_i(q_0, t, w) + q_i(q_0, t), \quad (2.22)$$

$$p_i^X(T_B) = 0 \quad (i = 1, \dots, n).$$

両方法ともにラグランジュ乗数 p_m^Z は次式で与えられる。

$$p_m^Z(k) = \sum_{j=1}^n p_j^X(t) W_{jm}^A \quad (m = 1, \dots, M). \quad (2.23)$$

学習則の導出は付録Aで行なう。

2.5 関数近似としての力学系学習

前節で述べた学習則はいずれもBP法を拡張したものであり、二乗誤差を最小化する勾配降下法に基づいている。

力学系の学習あるいは時系列の学習とは、本質的にはその時系列を生成する力学系の同定である。RNNで学習する場合は、対象とする力学系が有限次元 (n 次元)であることを仮定し、同次元を持ったRNNのダイナミクスを規定する関数 \hat{f} により学習対象の力学系を規定する関数 f を近似することである。したがって、最小化すべき誤差は本来、全空間での関数の誤差の上限值

$$\sup_{x \in R^n} \|\hat{f}(x, w) - f(x)\|^2 \quad (2.24)$$

である。しかし、現実には次のような問題があり、この関数そのものを評価する学習法は実行が困難である。

(1) 教師信号が与えられるのは通常、定常状態 (アトラクタ上) だけであるから、 n 次元空間全体での誤差は評価できない。

表2-1 学習方法の比較

	関数の近似	時系列の近似
離散系	関数誤差法	軌道誤差法
	$error = \tilde{f}(q(k), w) - f(q(k))$	$error = \tilde{f}^{k+1}(q(0), w) - f^{k+1}(q(0))$
連続系	速度誤差法	波形誤差法
	$error = \tilde{f}(q(t), w) - f(q(t))$	$error = x(q_0, t, w) - q(q_0, t)$
特徴	直接的、陽的	間接的、陰的

$$E = \frac{1}{2} \sum_{k=0}^{N_n-1} \|error\|^2 \text{ (離散系)}$$

$$E = \frac{1}{2} \int_0^{T_B} \|error\|^2 dt \text{ (連続系)}$$

(2) 連続系の場合、関数 f は教師信号の微分値であるので、観測困難であることが多い。

そこで、評価可能な誤差評価関数を何らかのかたちで近似することになり、その近似の仕方に応じて、前節で述べた個々の学習則が導出される(表2-1)。

離散系の関数誤差法、連続系の速度誤差法は誤差が関数 f で定義されているので、本来の関数近似の主旨に沿った評価関数になっている。ただし、速度誤差法の場合は教師信号の微分値は観測可能なものと仮定する。

連続系の波形誤差法は、教師信号の微分値が観測不可能な場合に、信号の波形そのものの誤差を評価する方法である。離散系の軌道誤差法も、関数の誤差よりもむしろ時系列そのものの誤差を評価している。あるいは、関数誤差法と速度誤差法が関数 f の誤差を直接評価する陽的な方法であるのに対し、軌道誤差法と波形誤差法は時系列の誤差を評価することにより、それを生成する関数を近似していくような陰的な手法ということもできる。

関数誤差法と速度誤差法は関数 f の誤差を評価しているが、上記の問題点(1)のため、全空間での最大値を評価することはできない。そこで、アトラクタ上に制限されたデータだけを用い、さらに上限値ノルムよりも微分係数の算出が容易な L^2

ノルムにより定義された評価関数

$$E_P(w) = \frac{1}{2} \int_{\mathcal{R}^n} \|\hat{f}(q, w) - f(q)\|^2 dP \quad (2.23)$$

を最小化することになる。ここで、 P は教師信号の不変測度^{35, 36)}を表わす。 P はアトラクタの部分集合でのみ非零の値を持つ。教師信号のエルゴード性を仮定すると、上記のルベーク積分は時間平均に一致する。時間平均はたとえば連続系の場合は

$$\lim_{T \rightarrow \infty} \frac{1}{2T} \int_0^T \|\hat{f}(q(t), w) - f(q(t))\|^2 dt \quad (2.26)$$

となる。教師信号が周期的な場合は上記の無限区間での積分は1周期区間でのものに置き換えることができるが、非周期的な信号を学習するときは、厳密には無限区間での積分が必要になる。また、実際には周期軌道を学習する場合でも、周期すなわち積分区間が長くなれば誤差評価関数の勾配(微分値)の計算が難しくなる。

そこで、関数誤差法と速度誤差法では、学習のための適当な周期 $N_B(T_B)$ を設定し、 E_1 や E_3 のように、この区間での総和(積分値)を評価している。教師信号が周期的な場合、 $N_B(T_B)$ をその周期(の整数倍)に一致させると、この総和(積分)は時間平均値と等価(周期の $N_B(T_B)$ 倍)である。特に離散系の関数誤差法では、有限個の周期点の上での入出力パターン $(x, f(x))$ の組をすべて提示し、その誤差の総和を評価する一括型のBP法に他ならない。この場合は、誤差評価関数は学習回数に陽には依存しないので最急降下法になる。したがって、学習係数の一次元最適値探索を行えば学習は局所的に収束することが保証されている³⁶⁾。

しかし、一般には誤差評価関数は学習回数に陽に依存するので、最急降下法ではない。関数誤差法・速度誤差法は複数のデータ点での誤差を評価する一括型BP法の形式をしているが、逐次型のBP法の拡張と考えることができる。RNNの学習則と逐次型BP法との関連については、確率的降下法の理論^{43, 44)}に基づき、次章で詳しく論じる。

2.6 連続力学系の学習則の分類

本論文で扱う主な学習則は前節で述べた。しかし、RNNの学習則は連続力学系の学習を対象としてこの他にも数多く提案されている。それらをすべて紹介することは不可能であるが、本節では、本論文で扱うものも含め連続力学系の学習則を異なった二つの観点から四つに分類し、簡単にまとめておく。

(1) 最適制御型学習法と適応制御型学習法²⁹⁾

速度誤差法と波形誤差法はともに誤差評価時間間隔 T_B を定め、その長さの区間の二乗積分誤差を評価関数としてこれを減少させることにより学習を実現していた。これらの方法は制御理論における最適制御手法と本質的に同じであることから最適制御型の学習法と呼ばれる。学習対象が連続力学系であっても、その学習則は次のような周期 T_B ごとにのみ値を持つ離散力学系(差分方程式)になるのが特徴である。

$$w((k+1)T_B) = w(kT_B) - \eta \nabla_w E(w(kT_B), k). \quad (2.27)$$

これに対して、積分誤差ではなく各時刻 t における二乗誤差(速度または波形)を減少させる連続時間型の最急降下法による学習法も提案されている。すなわち、次のようなパラメータ w に関する微分方程式によって学習則が記述される。

$$\frac{dw(t)}{dt} = -\eta \nabla_w E(w(t), t). \quad (2.28)$$

ここで、 $E(w, t)$ が時刻 t における誤差を表しており、速度誤差法では、

$$E(w, t) = \frac{1}{2} \| \dot{f}(q(t), w) - f(q(t)) \|^2. \quad (2.29)$$

波形誤差法では、

$$E(w, t) = \frac{1}{2} \| x(q_0, t, w) - q(q_0, t) \|^2 \quad (2.30)$$

である。この方法はやはり制御理論の手法との類似性から、適応制御型の学習則と呼ばれている。適応制御型の学習法については、その収束性の確率的な評価について次章で簡単に考察する。

(2) 随伴方程式法と感度方程式法²⁴⁾

また、前節の最適制御型の波形誤差法のアルゴリズムでは、ラグランジュ乗数を用いた随伴方程式によって誤差評価関数の勾配 $\nabla_w E$ を計算していたが、 $\partial x_i / \partial w_j$ に関する微分方程式(変分方程式)を解くことによって E を計算する方法もある。前者を随伴方程式法、後者を感度方程式法と呼ぶことにする。両者の相違は同一の学習則を実行するための計算法の相違であるから、結果は理論的には一致する。しかし、解くべき微分方程式の個数を比較すると、学習対象の系の次元が n のとき、随伴方程式法では $2n$ 個であるのに対し、感度方程式法では $n^2 + 1$ 個を要する²⁴⁾。逆に、随伴方程式法は微分方程式の2点境界値問題の形式になったため、

逆時間方向の求解を必要とする。このために、 T_B 時間の波形を記憶しておくメモリが必要になる。感度方程式法では、微分方程式の求解はすべて時間前向きであるので、このようなメモリは不要である。

以上のように両者には一長一短があるが、微分方程式の個数の増加に伴う計算量の増大に比べてメモリの負荷はわずかであるような問題が多く、現実には随伴方程式法が広く採用されている。本論文でも、随伴方程式法のみを用いる。

2.7 結言

本章では、三層階層型 RNN の構造とダイナミクス、および学習アルゴリズムの説明を行なった。

力学系の学習を関数近似として捉える観点から、陽的な手法である関数誤差法・速度誤差法、陰的な手法である軌道誤差法・波形誤差法の相違点を以下の章で明らかにしていく。

第3章 力学系の学習過程の確率的降下法の理論に基づく考察

3.1 緒言

本章では、前章で述べたRNNによる離散力学系および連続力学系の学習則を確率的降下法の理論に基づいて考察する。

2.5節で述べたように、前章で紹介した力学系の学習則はすべて逐次型BP法の拡張であり、二乗誤差を最小化する勾配降下法に基づいている。そして、誤差評価関数は一般には学習回数に陽に依存するので最急降下法ではない。

BP法を含む二乗誤差最小化の学習則が一般には最急降下法ではなく、甘利^{43,44)}の意味での確率的降下法になることが指摘されている³⁷⁾。甘利は、パターンの分布が未知であるような学習識別問題に関する一般論を展開し、ランダムに与えられる個々のパターンに基づき逐次的に識別系のパラメータを修正する学習方法を提唱した。この方法は、識別系が平均的に最適状態に漸近するような勾配降下法であることから、確率的降下法と呼ばれている。

本章では、前章で述べた学習則がすべて確率的降下法であり、平均的にはBP法と等価であることを示し、この結果に基づいて学習の平均速度の評価を行なう。

3.2 確率的降下法としての学習則の定式化

パラメータ w をもった関数 $\hat{f}(w, x)$ で記述されるシステムにより関数 $f(x)$ を近似する問題を考える。まず、1点 $x \in R^n$ での関数 \hat{f} と f の誤差を

$$l(x, w) = \frac{1}{2} \|\hat{f}(x, w) - f(x)\|^2 \quad (3.1)$$

と表わし、学習において R^n 上の集合 A に含まれる点が出現する確率測度を $P(A)$ として、誤差評価関数を次のように測度 P に関するルベーク積分で定義する。

$$E_P(w) = \int_{R^n} l(x, w) dP. \quad (3.2)$$

そこで、パラメータ w を修正して $E_P(w)$ を漸次小さくしていく一般的な学習則を考える。

学習が i 回行なわれたあとのパラメータの値を w_i とし、 i 回目の入力 x_i にもとづいて、これを w_{i+1} に修正するものとする。この修正規則を

$$w_{i+1} = w_i + \eta a(x_i, w_i) \quad (3.3)$$

とする。 $a(x, w)$ はパラメータ w の修正量を決定する、 w と同じ次元のベクトル量であり、学習関数と呼ばれる。また、 η は学習係数と呼ばれる正の定数である。

1 回の学習により w が δw だけ修正されるとする。これに対応して $E_P(w)$ は

$$\delta E_P(w) = \delta w^T \nabla_w E_P + \frac{1}{2} \delta w^T \nabla \nabla_w E_P \delta w + O(\|\delta w\|^3) \quad (3.4)$$

だけ変化する ($\nabla \nabla_w E_P$ は w に関する微分による E_P の Hesse 行列を表わす)。

δw は x_i と w_i に依存し、その期待値 (x についての平均値) は式 (3.3) より、

$$\overline{\delta w}(w) = \eta \int a(x, w) dP \quad (3.5)$$

となる。したがって δE_P の期待値は

$$\overline{\delta E_P}(w) = \overline{\delta w}^T \nabla_w E_P + \frac{1}{2} \eta^2 \text{trace}(Q \nabla \nabla_w E_P) + O(\eta^3) \quad (3.6)$$

と表わされる。ここで、

$$Q = \int a(x, w) \cdot a(x, w)^T dP \quad (3.7)$$

とおいた。さらに、 w を確率変数と考え、その確率密度関数を $r_i(w)$ と書く。 i 回の学習後の期待値は

$$\hat{E}_{P_i} = \int r_i(w) E_P(w) dw \quad (3.8)$$

と表わせ、1 回の学習による \hat{E}_{P_i} の変化 $\delta \hat{E}_{P_i} = \hat{E}_{P_{i+1}} - \hat{E}_{P_i}$ は

$$\delta \hat{E}_{P_i} = \int r_i(w) \overline{\delta E_P}(w) dw \quad (3.9)$$

となる。

以上の定式化は、パターン識別の場合の甘利による定式化とほとんど同じである。主な相違点は次の二点である。第一に、パターン識別の場合はパターンの出現確率を確率密度関数で表現しているが、本論文では確率測度によるルベーク積分を用いて確率密度関数を持たないようなより一般的な場合にも対応している。第二に、パターン識別の場合は x の出現確率密度関数 $p(x)$ を有限個のカテゴリの出現確率を関与させた形で定義しているが、本論文では関数近似を扱っているためカテゴリによる分類を考えていない。しかし、この二点を除けば本質的には両者は同じである。そこで、以下、甘利による「学習収束の定理」にもとづいて前節で述べた学習則の収束性を考察する。

3.3 離散力学系の学習

離散力学系の学習では、教師信号 x_i は式 (2.5) の軌道に沿って与えられる。教師信号がその不変測度に関してエルゴード的³⁵⁾であることは一般に仮定できる。したがって、 P として f の不変測度を採用するのが妥当である。すなわち、 R^n の任意の可測集合 A に対して、次式が成り立つとする³⁵⁾。

$$P(A) = P(f^{-1}(A)). \quad (3.10)$$

3.3.1 B P 法による学習

B P 法 ($N_p = 1$ の場合の関数誤差法、軌道誤差法) の場合は、式 (3.3) の学習関数 $\alpha(x, w)$ が次式で与えられる。

$$\alpha(x, w) = -\nabla_w l(x, w) = -\frac{1}{2} \nabla_w \|\hat{f}(x, w) - f(x)\|^2. \quad (3.11)$$

このとき式 (3.2) と (3.5) より、 w の修正量の期待値は

$$\overline{\delta w}(w) = -\eta \nabla_w E_p(w) \quad (3.12)$$

となる。式 (3.12) は w の修正方向が平均的には誤差 E_p を減少させる方向を向いていることを意味している。つまり η^2 のオーダーを無視すれば式 (3.9) と (3.12) より、

$$\delta \dot{E}_p = -\eta \int r_i(w) \|\nabla_w E_p(w)\|^2 dw \leq 0 \quad (3.13)$$

となって、 E_p は減少する。すなわち、B P 法は確率的降下法の典型例の一つである³⁷⁾。

3.3.2 関数誤差法による学習

学習関数 $\alpha(x, w)$ を式 (2.6) の E_1 を用いて

$$\alpha(x, w) = -\nabla_w E_1(w) \quad (3.14)$$

と定義する。このとき式 (2.6) と (3.1)、(3.5)、(3.14) より、

$$\overline{\delta w}(w) = -\eta \sum_{i=0}^{N_p-1} \int_{R^n} \nabla_w l(f^i(x), w) dP \quad (3.15)$$

となる。ここで、 P は f の不変測度であったから、 x の任意の可積分関数 $G(x)$ と任意の正整数 k について、

$$\int_{R^n} G(f^k(x)) dP = \int_{R^n} G(x) dP \quad (3.16)$$

が成り立つ⁵⁵⁾。そこで $G(x) = \nabla_w l(x, w)$ の場合を考えると、式(3.15)より、 w の修正量の期待値は次のようになる。

$$\overline{\delta w}(w) = -\eta N_B \int_{R^n} \nabla_w l(x, w) dP = -N_B \eta \nabla_w E_P(w). \quad (3.17)$$

式(3.17)はBP法の場合の式(3.12)で学習係数 η を N_B 倍したものにほかならない。したがって、1サイクルあたりの学習回数 N_B を大きくすることは、 $N_B = 1$ の場合のBP法において学習係数 η を大きくすることと確率的な意味で等価であると考える。

3.3.3 軌道誤差法による学習

$$\alpha(x, w) = -\nabla_w E_2(w) = -\frac{1}{2} \nabla_w \sum_{k=0}^{N_B-1} \|j^{k+1}(x, w) - f^{k+1}(x)\|^2 \quad (3.18)$$

とする。この場合は、

$$\hat{l}(N_B, x, w) = \frac{1}{2} \sum_{k=0}^{N_B-1} \|j^{k+1}(x, w) - f^{k+1}(x)\|^2 \quad (3.19)$$

として、誤差評価関数を、

$$\hat{E}_P(N_B, w) = \int_{R^n} \hat{l}(N_B, x, w) dP \quad (3.20)$$

とした場合の確率的降下法になっている。また、

$$\hat{E}_P(N_B, w) \geq E_P(w) \quad (3.21)$$

であるから、 E_P も \hat{E}_P とともに減少することが期待される。

3.4 連続力学系の学習

連続力学系の学習では、教師信号 x_i は式(2.15)の解軌道 $q(t)$ に沿って与えられる。すなわち、

$$x_i = q(t + T_B i)$$

である。したがって、 P は $q(t)$ の不変測度と考える。連続系の場合も教師信号のエルゴード性は仮定しておく^{*}。教師信号が混合的ならば、その $\text{time-}T_B$ map も混合的、したがってエルゴード的だから教師信号の時間平均は不変測度に関する空間平均に一致する。

すなわち、 R^n の任意の可測集合 A に対して、次式が成り立つものと仮定する。

$$P(A) = P(q(A, -t)) \quad (3.22)$$

ただし、

$$q(A, -t) = \{ q_0 \in R^n \mid q(q_0, t) \in A \}$$

である。

3.4.1 速度誤差法による学習

学習関数を

$$a(x, w) = -\nabla_w E_3(w). \quad (3.23)$$

とする。式(2.16)と(3.23)より、

$$\begin{aligned} a(x, w) &= -\frac{1}{2} \nabla_w \int_0^{T_B} \|\hat{f}(q(x, t), w) - f(q(x, t))\|^2 dt \\ &= -\int_0^{T_B} \nabla_w l(q(x, t), w) dt. \end{aligned} \quad (3.24)$$

したがって、式(3.5)と(3.24)より、

$$\bar{\delta w} = -\eta \int_0^{T_B} \int_{R^n} \nabla_w l(q(x, t), w) dP dt. \quad (3.25)$$

仮定より、 P は $q(x, t)$ の不変測度であるから、 R^n 上の任意の可積分関数 F と任意の $t \in R$ に対して、

$$\int_{R^n} F(q(x, t)) dP = \int_{R^n} F(x) dP$$

が成り立つ。よって、

$$\int_{R^n} \nabla_w l(q(x, t), w) dP = \int_{R^n} \nabla_w l(x, w) dP. \quad (3.26)$$

^{*} 教師信号が周期的あるいは概周期的（あるいは、このような集合とカオスの直積、たとえば強制系のカオス）で T_B がその周期成分の逆数と有理数比を持つ場合を除く。このような場合の生じる確率は零である。

したがって、 w の修正量の期待値は、

$$\begin{aligned}\bar{\delta w} &= -T_B \eta \int_{R^*} \nabla_w l(x, w) dP \\ &= -T_B \eta \nabla_w E_P l(w).\end{aligned}\quad (3.27)$$

となる。これは学習係数が $T_B \eta$ の B P 法 (式 (3.12)) のそれに等しい。

3.4.2 波形誤差法による学習

学習関数は次のように定義される。

$$a(x, w) = -\nabla_w E_4(w).\quad (3.28)$$

また、

$$\begin{aligned}q(x, t) &= x + \dot{q}(x, 0)t + O(t^2), \\ x(x, t, w) &= x + \dot{q}(x, 0, w)t + O(t^2)\end{aligned}\quad (3.29)$$

であるから、

$$\|x(x, t, w) - q(x, t)\|^2 = \|\dot{f}(x, w) - f(x)\|^2 t^2 + O(t^3)\quad (3.30)$$

となる。よって、式 (2.17) と (3.28)、(3.30) より、

$$a(x, w) = -\frac{1}{3} T_B^3 \nabla_w l(x, w) + O(T_B^4).\quad (3.31)$$

したがって、

$$\bar{\delta w} = -\frac{1}{3} T_B^3 \eta \nabla_w E_P l + O(\eta T_B^4).\quad (3.32)$$

T_B と η が十分小さいとき、これは B P 法において学習係数を $(1/3)T_B^3 \eta$ としたものとみなすことができる。

学習係数 η が十分小さい場合、B P 法の学習に要する時間 T_C は η に反比例するので、速度誤差法については、

$$T_C \propto (T_B \eta)^{-1}.\quad (3.33)$$

波形誤差法については、

$$T_C \propto (T_B^3 \eta)^{-1}\quad (3.34)$$

と見積もることができる。

3.5 適応制御型学習則の収束性

第2章で述べた適応制御型の学習則についても、その収束性を確率的降下法の理論に基づいて議論できる。甘利による議論は基本的に離散時間での学習である。ここでは、これを連続時間に拡張することを試みる。連続時間系の勾配降下法は離散系のそれより議論において近似が少なくて済む。また、ここではパラメータ空間のコンパクト性を仮定することにより議論を厳密化している。

適応制御型の学習則は次の微分方程式で記述される連続時間型の勾配降下法であった。

$$\frac{dw(t)}{dt} = -\eta \nabla_w E(w(t), t). \quad (3.35)$$

以下、誤差評価関数として

$$E(w, t) = \frac{1}{2} \|\hat{f}(q(t), w) - f(q(t))\|^2 \quad (3.36)$$

を用いる速度誤差法を考える。

まず、式(3.2)と(3.36)より

$$\frac{dE_P(w(t))}{dt} = \nabla_w E_P(w) \cdot \frac{dw(t)}{dt}. \quad (3.37)$$

式(3.35)より、 $dw(t)/dt$ は $q(t)$ に依存しているから、 $dE_P(w(t))/dt$ も $q(t)$ に依存する。そこで、式(3.37)を q に関して平均すると次式のようになる。

$$\begin{aligned} \frac{dE_P(w(t))}{dt} &= \nabla_w E_P(w(t)) \int \left(-\frac{\eta}{2}\right) \nabla_w \|\hat{f}(q, w(t)) - f(q)\|^2 dP \\ &= -\eta \|\nabla_w E_P(w(t))\|^2 \leq 0. \end{aligned} \quad (3.38)$$

よって、学習は $E_P(w)$ を平均的に小さくするように進行する。

式(3.38)は最適制御型学習則の式(3.13)に相当するが、(3.13)とは異なり $O(\eta^2)$ の近似は入っていない。この相違は、離散型の勾配降下法が本来は連続型のその近似であることに起因する。

以下、甘利による学習収束の定理に倣い、パラメータ空間がコンパクトという仮定の元で、適応制御型の学習が最適状態に確率1で収束することを証明する。

まず、学習の最適状態の集合を

$$W_{OP} = \{w \mid \nabla_w E_P(w) = 0\}$$

と定義する ($E_P(w) = 0$ とは限らない)。

さて、 $w(t)$ は時刻 0 から t までの教師信号の値 $q(\tau)$ ($\tau \in [0, t]$) に依存している。したがって、 $q(t)$ を確率過程と考えると、 $w(t)$ も確率変数である。そこで、時刻 t におけるその確率測度を $R^{(t)}$ で表わし、 $E_P(w)$ の期待値を

$$\hat{E}(t) = \int E_P(w) dR^{(t)} \quad (3.39)$$

とする。

さて、

$$\hat{E}(t + \Delta t) = \int E_P(w) dR^{(t+\Delta t)} = \int E_P(w + \overline{\Delta w(t)}) dR^{(t)} \quad (3.40)$$

であり、

$$E_P(w + \overline{\Delta w(t)}) = E_P(w) + \frac{dE_P(w)}{dt} \cdot \Delta t + O(\|\Delta t\|^2) \quad (3.41)$$

である。したがって、

$$\begin{aligned} \frac{d\hat{E}(t)}{dt} &= \lim_{\Delta t \rightarrow 0} \frac{\hat{E}(t + \Delta t) - \hat{E}(t)}{\Delta t} \\ &= \int \frac{dE_P(w)}{dt} dR^{(t)} \end{aligned} \quad (3.42)$$

となる。

以上の準備のもとに、適応制御型学習の収束定理を述べる。

[定理 3-1]

$$M'_\mu(\eta) = R^{(\eta)}(\{\|w(t) - W_{OP}\| \geq \mu\}),$$

$$M_\mu(\eta) = \lim_{\eta \rightarrow \infty} M'_\mu(\eta)$$

とすると、もし、パラメータ w の全体の空間がコンパクトであれば、任意の $\eta > 0$ 、任意の $\mu > 0$ に関して、

$$M_\mu(\eta) = 0$$

が成立する。

(証明)

$$U_\mu = \{ w \mid \|w - W_{OP}\| \geq \mu \},$$

$$U'_\lambda = \{ w \mid \|\nabla_w E_P(w)\|^2 \geq \lambda \}$$

とすると、任意の $\mu > 0$ に対して、 $\lambda > 0$ が存在して、 $U_1 \cap U_\mu$ となる。なぜならば、 U_μ はコンパクト空間の閉集合だからコンパクトである。よって、 U_μ 上で $\|\nabla_{\mathbf{w}} E_P(\mathbf{w})\|^2$ の最小値が存在する。この値を λ とすればよい。

さて、

$$M'_\mu(\eta) = \int_{U_\mu} dR^{(t)}$$

と書けるから、

$$\begin{aligned} \int \|\nabla_{\mathbf{w}} E_P(\mathbf{w})\|^2 dR^{(t)} &\geq \int_{U_\mu} \|\nabla_{\mathbf{w}} E_P(\mathbf{w})\|^2 dR^{(t)} \\ &\geq \lambda \int_{U_\mu} dR^{(t)} \geq \lambda \int_{U_\mu} dR^{(t)} = \lambda M'_\mu(\eta). \end{aligned} \quad (3.43)$$

また、式(3.38)と(3.42)より、

$$\frac{d\hat{E}(t)}{dt} = -\eta \int \|\nabla_{\mathbf{w}} E_P(\mathbf{w})\|^2 dR^{(t)}. \quad (3.44)$$

よって、

$$\frac{d\hat{E}(t)}{dt} \leq -\eta \lambda M'_\mu(\eta). \quad (3.45)$$

したがって、

$$\begin{aligned} \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \frac{d\hat{E}(t)}{dt} &= \lim_{T \rightarrow \infty} \frac{1}{T} [\hat{E}(T) - \hat{E}(0)] \\ &\leq -\eta \lambda M'_\mu(\eta) \end{aligned} \quad (3.46)$$

となる。今、 \mathbf{w} の全体はコンパクトだから、 $E_P(\mathbf{w})$ は有界なので、上式左辺は 0 になる。よって、 $\eta > 0$ 、 $\lambda > 0$ より、 $M'_\mu(\eta) = 0$ となる。□

3.6 結言

本章では、RNNによる力学系の学習過程を確率的降下法の理論を用いて解析した。

まず、第2章で紹介した最適制御型の学習則がすべて甘利の意味での確率的降下法として定式化できることを示した。

また、離散力学系の関数誤差法と連続力学系の速度誤差法は、教師信号の不変測度に関して平均すると、ともにBP法と等価であることを証明し、学習速度を学習係数と誤差評価時間とから評価する式を導いた。

さらに、連続力学系に対する適応制御型学習則が学習係数の値に依存せず、最適集合に確率1で収束することを証明した。

第4章 力学系の学習実験

4.1 緒言

離散力学系および連続力学系の学習の計算機実験を行なった。実験の目的は、次の3点である。

- (1) 離散系を対象とした関数誤差法と軌道誤差法、連続系を対象とした速度誤差法と波形誤差法による時系列の学習過程を観測し、前章で考察した確率的降下法の理論に基づく学習速度の推定の妥当性を検証すること。
- (2) 力学系の学習を関数の近似として捉える観点から、陽的な手法である関数誤差法、速度誤差法と陰的な手法である軌道誤差法、波形誤差法との相違点を明確にすること。
- (3) 教師信号が有限時間のみ提示できる場合とその分布密度を変化させた場合の学習結果を、教師信号を時系列どおりに任意の時間長にわたって提示する通常の方法による学習結果と比較し、教師信号の提示方法が学習に及ぼす影響を考察すること。

本章では、まず、離散力学系の学習実験結果を述べ、上記の点に関して検討を行なう。次に、離散系のカオス時系列の学習において観測された学習の加速現象について述べ、その原因について詳しく考察する。次に、教師信号の提示方法が学習に及ぼす影響を、離散力学系の学習を対象として考察する。最後に、連続力学系の学習実験結果を述べる。

4.2 離散力学系の学習

Cubic map と呼ばれる次式(4.1)で記述される1次元離散力学系のアトラクタを関数誤差法および軌道誤差法で学習し、両者の比較を行なう。

$$q(k+1) = \mu q(k)(1 - q^2(k)) \quad (q \in \mathbb{R}). \quad (4.1)$$

学習の対象は図4-1に示した3種類のアトラクタとする。図4-1(a)は4周期軌道、(b)と(c)はカオスである。(a)、(b)、(c)の各アトラクタのリアプノフ指数^{57, 58)}はそれぞれ-0.177、0.403、0.806である。

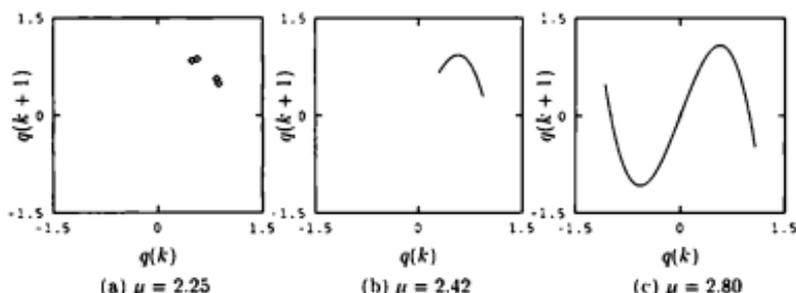


図4.1 1次元離散力学系 $q(k+1) = \mu q(k)(1 - q^2(k))$ のアトラクタ

4.2.1 学習曲線と学習に要する時間の定義

離散力学系の学習の進行状況を把握するため、教師信号とRNNとの誤差の学習にもなう変化を測定する。

勾配降下法に基づく学習アルゴリズムはすべて、ある誤差評価関数を定義し、これを減少させることにより学習を進めている。したがって、この誤差評価関数の値の変化によって学習の進行の度合を把握するのが自然である。しかし、各アルゴリズムによって誤差評価関数は異なるので、アルゴリズムの性能を比較するためには、個々のアルゴリズムに依存しない評価基準を定義しておく必要がある。そこで、第2章の式(2.5)で記述される一般の離散力学系の学習の評価基準を以下のように定める。

第2章で述べたように、力学系の学習とは本質的には関数の近似であるから、第3章の式(3.2)で定義される $E_P(w)$ で学習の進行状況を測定するのが妥当である。しかし、不変測度 P を求めて $E_P(w)$ を厳密に計算することは困難なので、ここでは学習する力学系のアトラクタ上の N_S 個のサンプリング点 $\mathbf{x}(j)$ ($j = 1, \dots, N_S$) での二乗関数誤差

$$E_S = \sum_{j=1}^{N_S} \|f(\mathbf{x}(j)) - \hat{f}(\mathbf{x}(j), \mathbf{w})\|^2 \quad (4.2)$$

で $E_p(w)$ を近似する。サンプリング点 $\{x(1), \dots, x(N_S)\}$ は式(2.5)の力学系によって生成されたアトラクタの中から一様にサンプリングするので、式(4.2)は式(3.2)で P を f の不変測度としたものを近似していることになる。 E_S は学習とともに変化していくので、学習回数 i の関数 $E_S(i)$ となり、横軸に学習回数、縦軸に $E_S(i)$ をとったグラフを本論文では学習曲線と呼ぶことにする。学習曲線の一例を図4-2に示す。

また、学習に要する時間 T_C を学習初期の段階で誤差が急激に変化する時刻で定義する。すなわち、図4-2のように誤差 E_S が一定のしきい値 E_C を切る時刻を T_C とする。

4.2.2 関数誤差法による学習過程

まず、関数誤差法を用いた学習で、学習係数 η と1学習サイクルでの学習回数 N_B を変化させた場合の、学習に要する時間の変化を調べる。誤差評価関数 E_S を測定するためのアトラクタ上のサンプル数は $N_S = 100$ とした。また、教師信号を与える Cubic map (4.1) の初期値は、 $q(0) = 0.1$ から出発し、10万回の写像を繰り返して過渡状態を取り除いた後のものを用いた。また、結合荷重の初期値はすべて $-0.001 \sim 0.001$ の範囲の値をとる一様乱数により与えた。

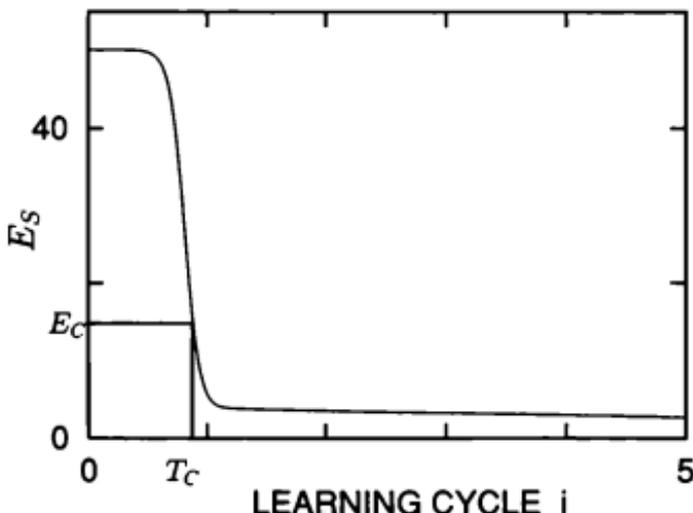


図4-2 学習曲線の例と E_C 、 T_C の定義

図4-1に示した三つのアトラクタについての関数誤差法による学習曲線を図4-3に示す。ここでは各アトラクタについて、

(i) $\eta = 0.01, N_B = 1$, (ii) $\eta = 0.04, N_B = 1$, (iii) $\eta = 0.01, N_B = 4$

の三とおりの条件で学習を行なった。シグモイドユニット数は $M = 3$ とした。(i) と (ii) は $N_B = 1$ なので B P 法に一致する。

図4-3を見ると、各アトラクタとも (ii), (iii) のほうが (i) よりも学習が速く、(ii) と (iii) には学習初期の曲線で見える限り大差がない。すなわち学習係数 η を4倍することと、 N_B を4倍することとはほぼ等価である。

一方、学習が進んだ段階での学習曲線の様子を見ると、4周期軌道(a)とカオス(b)については(ii)と(iii)とで大差はないが、カオス(c)については、(ii)のほうが(iii)よりも誤差の減衰が速くなっている。

学習初期の段階での学習速度を詳しく見るために、 η と N_B と学習に要する時間 T_C との関係を図4-4に示す。図4-4(i)は学習係数 η を一定にした場合の N_B と T_C の関係、図4-4(ii)は N_B を一定にした場合の η と T_C の関係を示しており、ともに両対数軸で表現している。誤差のしきい値 E_C は、(a)については5.0、(b)と(c)については10.0とした。この図より、 η と N_B がともに小さい範囲では、

$$T_C \propto \frac{1}{N_B \cdot \eta} \quad (4.3)$$

の関係が成立していることがわかる。これは、関数誤差法による学習は学習係数が $N_B \cdot \eta$ の B P 法と等価であることを意味し、第3章で述べた確率的降下法の理論にもとづくパラメータ修正量の期待値の評価式(3.17)を裏付ける結果であると言える。リアブノフ指数の大きいカオス(c)の場合、 η と N_B が大きいところではこの関係が少し乱れている。これは式(3.17)に基づき B P 法として換算した等価的な学習係数 $N_B \cdot \eta$ が大きくなって学習が不安定になるためと考えられる。

4.2.3 軌道誤差法による学習過程

軌道誤差法についても図4-1に示した三つのアトラクタの学習を行ない、学習曲線を図4-5に、 η と N_B と T_C との関係を図4-6に示した。図4-5でのパラメータの設定は関数誤差法と同じく、

(i) $\eta = 0.01, N_B = 1$, (ii) $\eta = 0.04, N_B = 1$, (iii) $\eta = 0.01, N_B = 4$

であり、シグモイドユニット数 M と E_C の設定も関数誤差法の場合と同じとした。まず、4周期解(a)とカオス(b)については、図4-5、図4-6から、関数誤差法と同様に

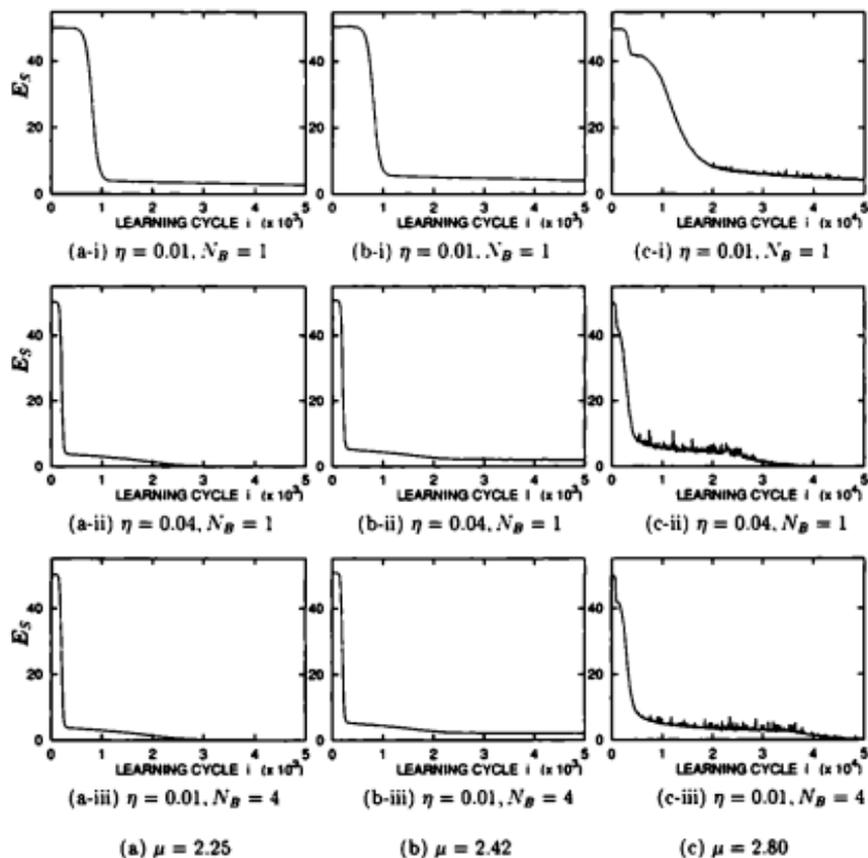


図 4-3 関数誤差法による学習の学習曲線

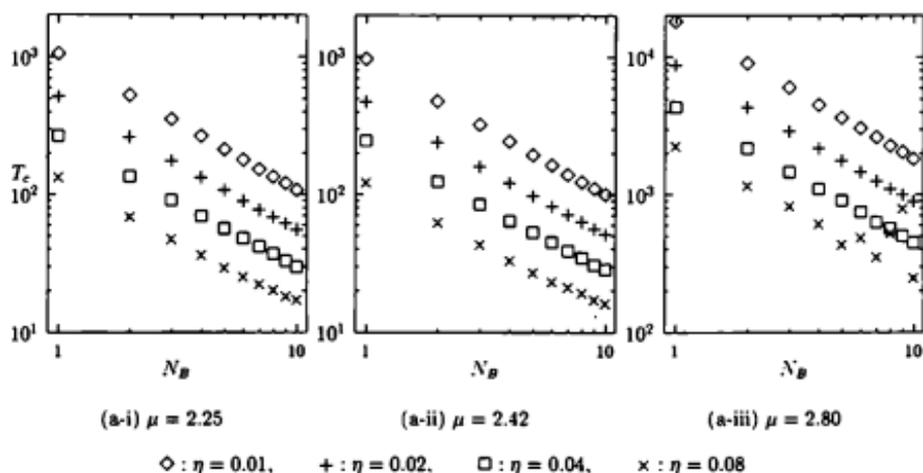


図 4-4 (i) 関数誤差法における学習パラメータと学習時間との関係 (N_B と T_C との関係)

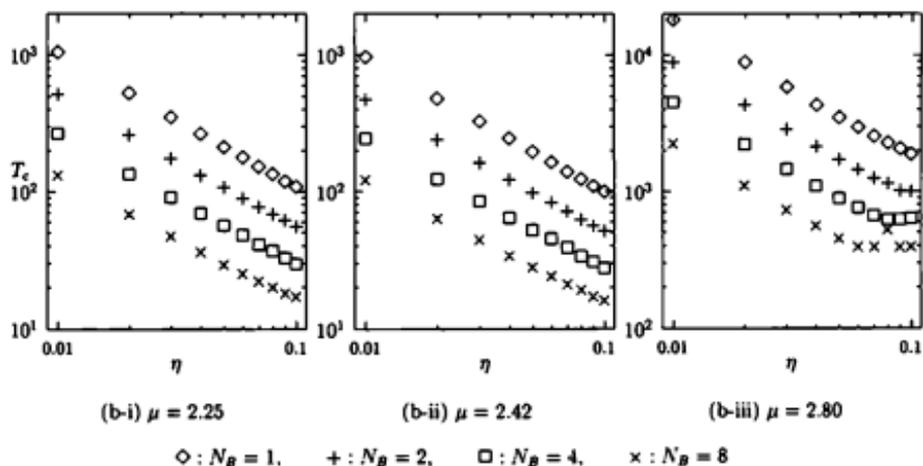


図 4-4 (ii) 関数誤差法における学習パラメータと学習時間との関係 (η と T_C との関係)

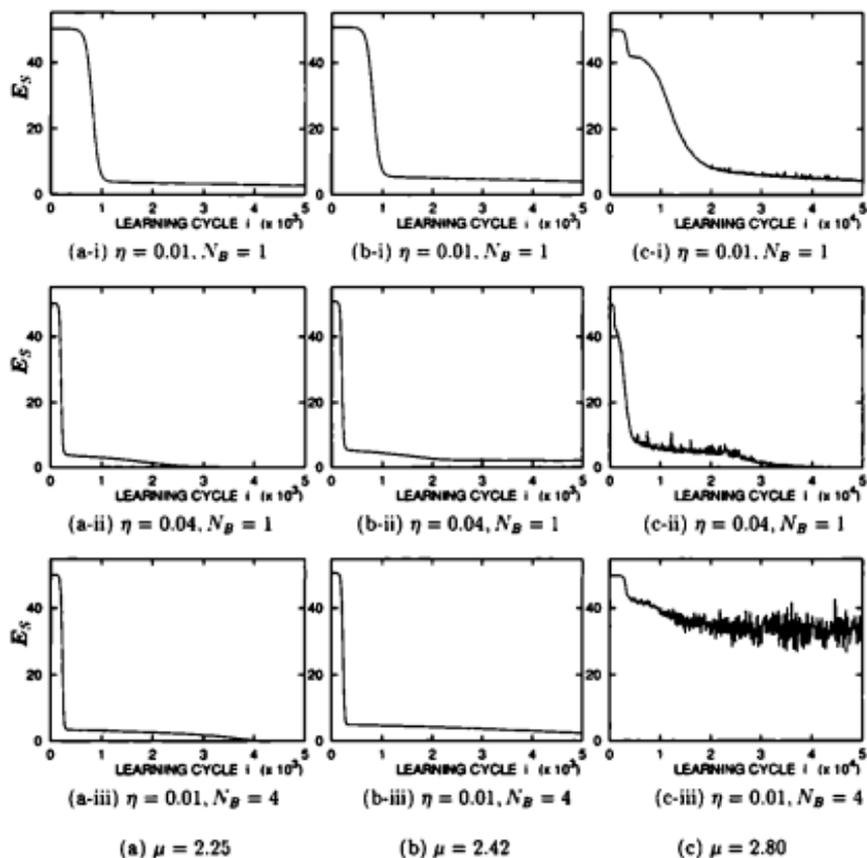


図 4-5 軌道誤差法による学習の学習曲線

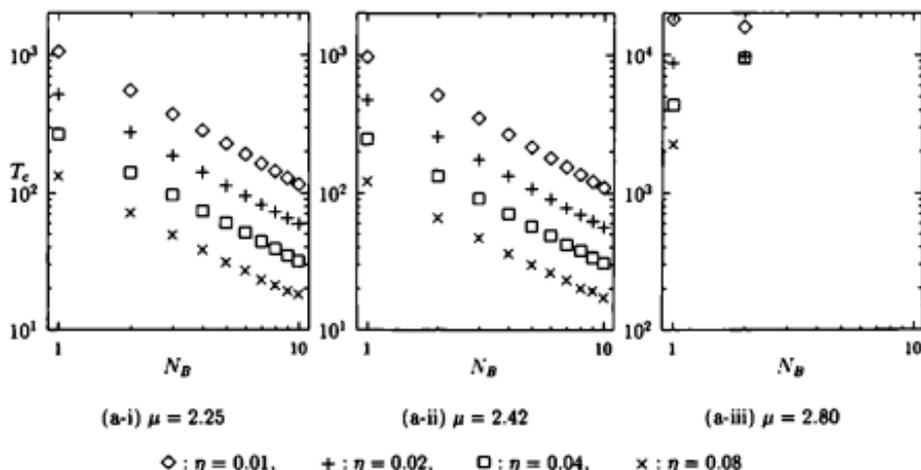


図 4-6 (i) 軌道誤差法における学習パラメータと学習時間との関係 (N_B と T_c との関係)

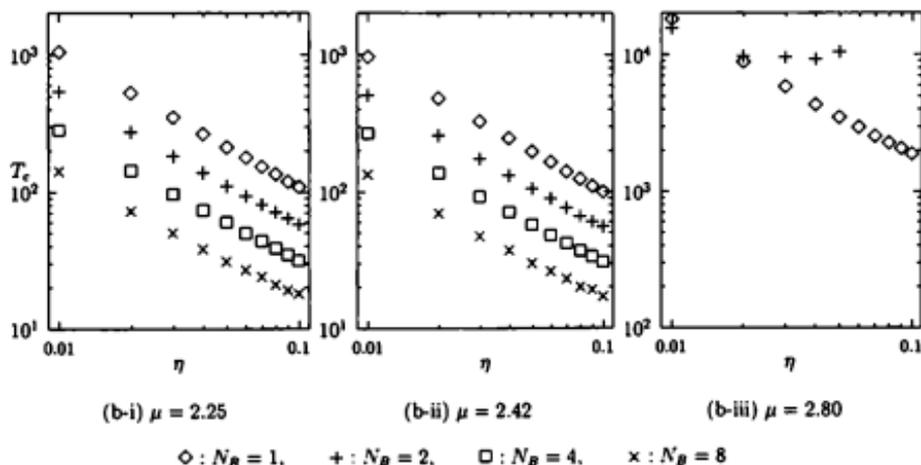


図 4-6 (ii) 軌道誤差法における学習パラメータと学習時間との関係 (η と T_c との関係)

式(4.3)の関係がほぼ成立していることがわかる。

しかし、カオス(c)の場合は、 N_B を大きくしても学習初期段階の学習速度はあまり増加せず、むしろ学習曲線が大きく振動し、学習の不安定化が激しくなる。図4-6では、 $N_B > 3$ の場合は誤差 E_S が E_C より小さくならず、 T_C がほとんど算出不能であった。これは、リアプノフ指数の大きいカオスの場合には本質的に短期間の軌道しか追跡できないため、1サイクルあたりの学習回数 N_B を多くしても、BP法にノイズを加えただけの効果しかないことを意味していると思われる^{23,24)}。

4.2.4 軌道誤差法における学習の加速

リアプノフ指数の大きいカオスを軌道誤差法で学習しても、学習初期の速度は増加しないことがわかった。しかし、軌道誤差法を用いた学習で、 N_B を適当な値に選ぶと、BP法や関数誤差法では見られない、学習の急激な加速が観測された。その一例を図4-7に示す。図4-7は、(a)BP法($N_B=1$)、(b)軌道誤差法($N_B=2$)、(c)関数誤差法($N_B=2$)、(d)確率的BP法($N_B=1$)の四とおりの方法で図4-1(c)のカオスを学習した結果の(i)学習曲線と(ii)RNNの出力の分岐図、および(iii)リアプノフ指数の変化を示している。学習係数 η はすべて0.01とした。

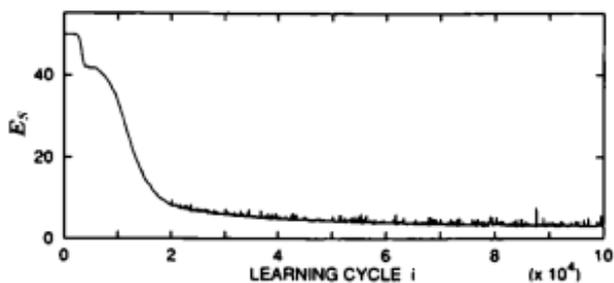
図4-7(ii)の分岐図は、 i 回の学習の後の結合荷重 $w(i)$ の値を用いて式(2.1)に従ってRNNを駆動した場合の出力 $x(k)$ のアトラクタの学習回数 i に対する変化を示している。結合荷重 w は系(2.1)のシステムパラメータであるから、学習にともなうパラメータ w の変化によるアトラクタの分岐現象を表示していることになる。同様に、図4-7(iii)には、 $w(i)$ の値を用いて駆動したRNNのアトラクタのリアプノフ指数の学習回数 i に対する変化を示している。

ここで、確率的BP法とは、BP法のパラメータ更新則にランダムノイズ項を付加したものであり、シミュレーテッドアニーリング法²⁹⁾と類似の効果により学習性能が向上することが報告されている³⁰⁾。具体的には、第2章の式(2.8)の更新則の右辺に、平均が0、分散が

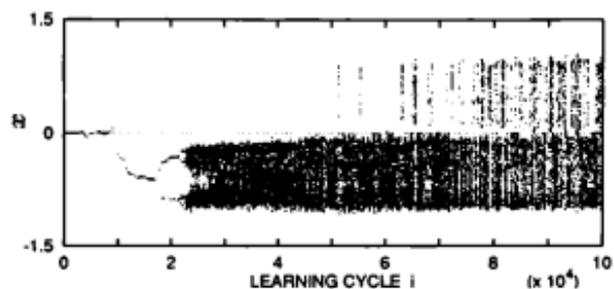
$$V = \frac{\beta^2}{\log(1+i)}$$

の白色ガウス雑音を付加する。ここで、 β は定数、 i は学習サイクル数である。図4-7(d)は $\beta=0.005$ の場合の結果を示している。

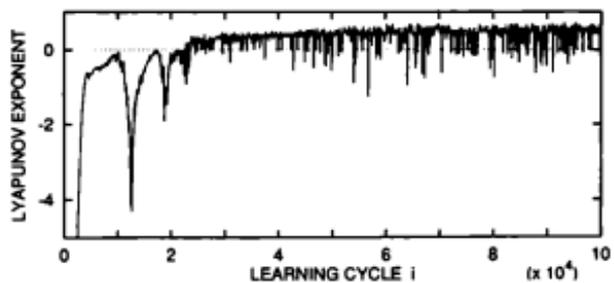
図4-7の(a)と(b)を比較すると、(b)の軌道誤差法による学習曲線は、学習初期の段階では N_B を2倍にした効果が現れず、(a)のBP法とほぼ同じ形状になっているが、学習サイクルが20,000あたりから曲線が乱れ始め、30,000付近で誤差が急激に



(i) 学習曲線



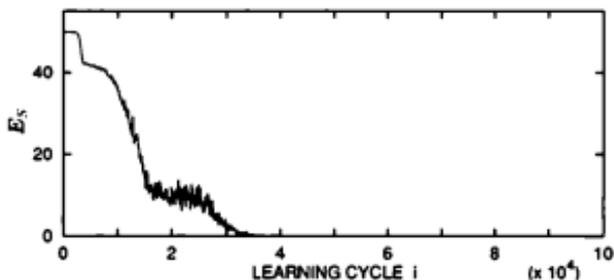
(ii) 分岐図



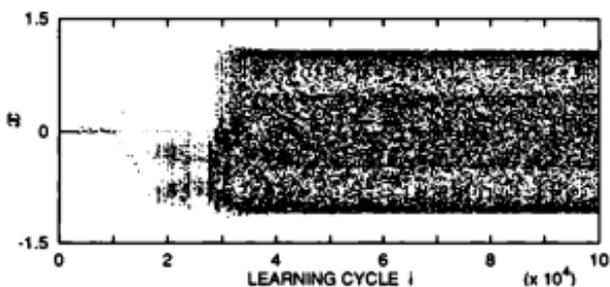
(iii) リアプノフ指数の変化

(a) B P 法 ($N_B = 1, \eta = 0.01$)

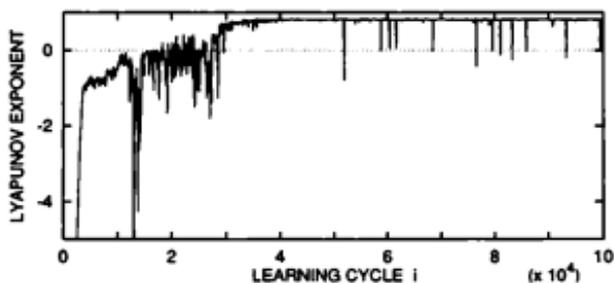
図 4-7 各種の方法によるカオスの学習過程 ($\mu = 2.80$)



(i) 学習曲線



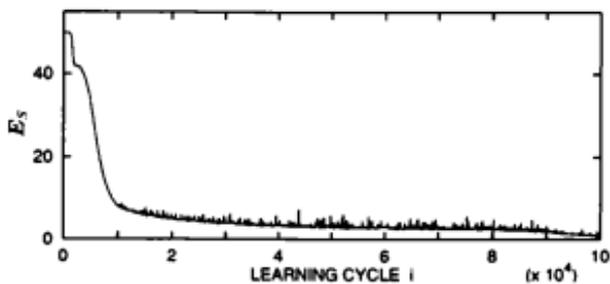
(ii) 分岐図



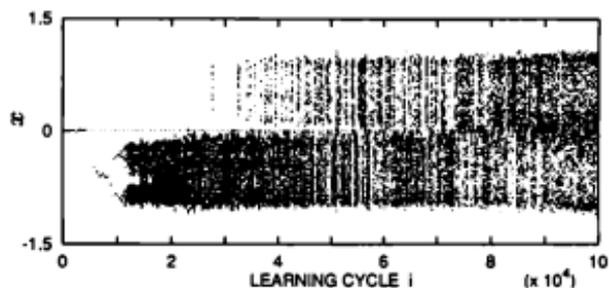
(iii) リアブノフ指数の変化

(b) 軌道誤差法 ($N_B = 2, \eta = 0.01$)

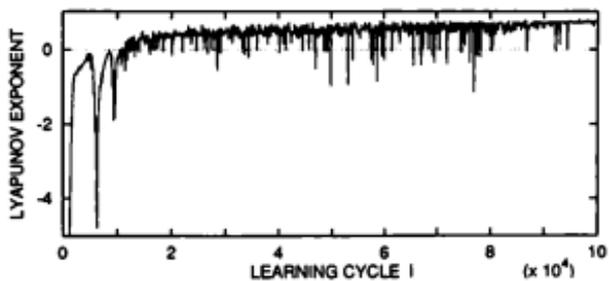
図 4-7 各種の方法によるカオスの学習過程 (続き)



(i) 学習曲線



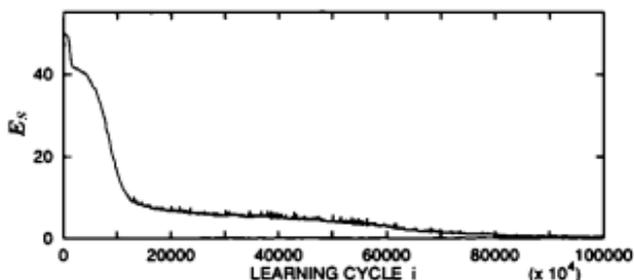
(ii) 分岐図



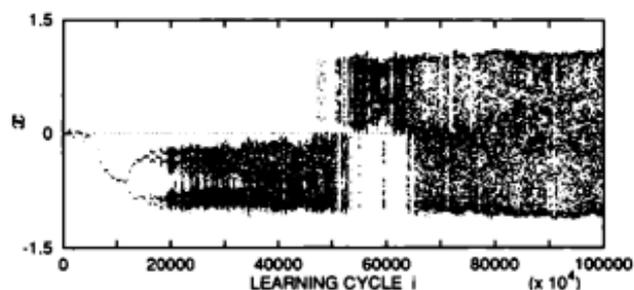
(iii) リアプノフ指数の変化

(c) 関数誤差法 ($N_B = 2, \eta = 0.01$)

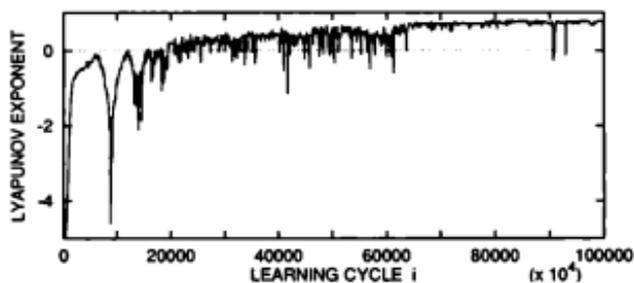
図 4-7 各種の方法によるカオスの学習過程 (続き)



(i) 学習曲線



(ii) 分岐図



(iii) リアプノフ指数の変化

(d) 確率的BP法 ($N_B = 1, \eta = 0.01, \beta = 0.005$)

図 4-7 各種の方法によるカオスの学習過程 (続き)

減少していることがわかる。(c)の関数誤差法では、前項の結果と同様、学習初期の速度が大きく(約2倍に)なっているが、軌道誤差法のような学習の急激な進行はみられない。

分岐図およびリアプノフ指数の変化をみると、軌道誤差法の場合、誤差が急激に減少する前の時点ですでにRNNの出力はカオスになっており、カオスの不安定性による「揺らぎ」が確率的BP法におけるノイズの役割を果たしているものと考えられる。さらに、(d)の確率的BP法と比較すると、確率的BP法でもノイズの効果によって学習が速くなっているが、軌道誤差法に比べると学習の進行は遅く、かつ不安定である。これは次のように解釈できる。すなわち、確率的BP法における揺らぎは人工的に付加したノイズによるものであり、学習すべきダイナミクスとはもともと無関係である。これに対し、軌道誤差法における揺らぎは学習対象のダイナミクスに起因するカオスによるものであるから、学習の進行とともにほどよく付加される。さらに、確率的BP法のノイズとは違って、時間的に減衰させる必要がない。このような揺らぎの特徴が学習をうまく加速しているものと考えられる。

次項では、ここで観測されたカオス学習における加速現象の発生メカニズムについて詳細に検討する。

4.2.5 学習の加速の発生メカニズム

ここでは、上に述べたカオスの学習における加速現象が次のような原因により発生していることを明らかにする。

- (1) カオスの持つ乱数的性質が逐次型の更新則において揺らぎの効果を与える。
- (2) 軌道誤差法において複数ステップの誤差を評価していることが(1)の揺らぎを拡大する。
- (3) この揺らぎがシグモイド関数のバイアス項に相当する結合荷重 W^D の変化を活性化し、関数の近似の精度を向上させる。

以下、このことを順を追って説明する。

まず、次の四つの方法での系(4.1)の学習時間を、教師信号のシステムパラメータ μ を変えつつ比較し、どのようなパラメータ値のアトラクタの学習において加速が生じるかを調べる。

- (a) BP法 ($\eta = 0.01, N_B = 1$)
- (b) 関数誤差法 I ($\eta = 0.01, N_B = 2$)

(c) 関数誤差法 2 ($\eta = 1.0 \times 10^{-4}, N_B = 100$)

(d) 軌道誤差法 ($\eta = 0.01, N_B = 2$)

B P 法と関数誤差法 2 とは、 $N_B \cdot \eta$ の値が等しいので ($N_B \cdot \eta = 0.01$)、第 3 章で述べた確率的降下法の理論よれば、両者は学習係数が $N_B \cdot \eta$ の B P 法として等価であるとみなすことができる。また、関数誤差法 2 の N_B を 100 という大きな値に設定したのは、これが一括型の B P 法 (最急降下法) の近似になることを意図したためである。

また、 E_S が学習初期時刻 (100 サイクル学習後) の誤差の 20%、1% になった時刻をそれぞれ、初期学習時間 T_1 、学習終了時間 T_2 と定義し、これらを学習速度の指標とする。

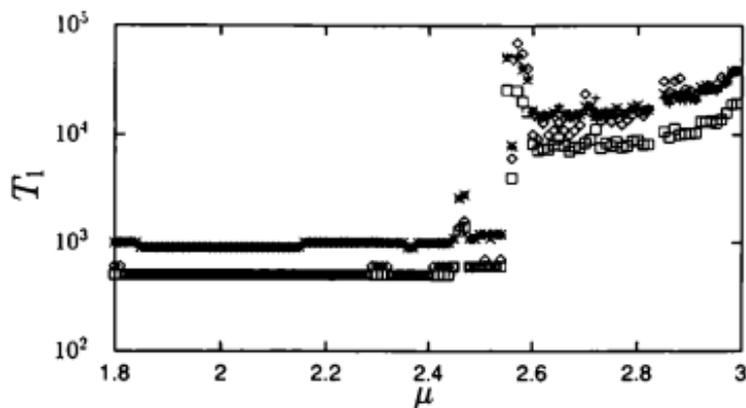
図 4-8 (a) および (b) はパラメータ μ を 1.80 から 3.00 まで変化させた場合の初期学習時間と学習終了時間の変化をそれぞれ示している。また、図 4-9 と図 4-10 には、同じ範囲での μ の変化に対する教師アトラクタの分岐ダイアグラムとリアブノフ指数の変化をそれぞれ示した。

まず、図 4-9、4-10 により、教師アトラクタの分岐とリアブノフ指数の変化の概要を説明しておく。

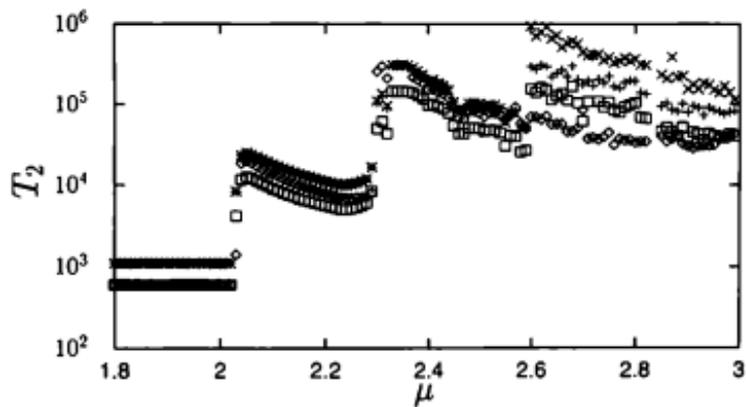
図 4-9 をみると、 $\mu = 1.80$ ではアトラクタは安定な 1 周期点であり、 μ を増加させると、 $\mu = 2.0$ で 1 周期点が不安定化し安定な 2 周期点が生じる、いわゆる周期倍分岐が起こることがわかる。 μ をさらに増加させると、周期は 4、8、16、… と増加し、 $\mu = 2.31$ 付近でアトラクタはカオスになる。カオスアトラクタのサイズは μ の増加とともに拡大し、途中「窓」と呼ばれる周期アトラクタの状態を経て、 $\mu = 2.60$ 付近で急激に拡大し、原点对称なアトラクタになる。これは、カオスアトラクタが不安定不動点 $x = 0$ と衝突し、 $x < 0$ の領域に存在するが図には示していないもう一つのカオスアトラクタと融合したことにより生じた現象で、「クライシス (crisis)」と呼ばれている。クライシスが生じる正確な値は解析的に算出でき、 $\mu = 3\sqrt{3}/2$ である。クライシスの後もアトラクタは拡大し続け、 $\mu = 3.0$ を超えると発散する。

また、図 4-10 より、リアブノフ指数はアトラクタが周期点の間は負の値であり、 $\mu = 2.31$ 付近でのカオスへの変化にともない、正になることがわかる。カオスアトラクタのリアブノフ指数は μ にたいしてほぼ単調に増加している。リアブノフ指数が負になる点はいくつかあるが、これは前述の窓に対応する。

次に、図 4-8 (a) で初期学習時間 T_1 を検討する。 $\mu = 1.80 \sim 2.54$ のアトラクタが周期点からクライシス前のカオスの範囲では、四つの方法とも T_1 の値はほぼ一定で



(a) 初期学習時間



(b) 学習終了時間

図 4-8 教師信号のパラメータ μ と学習時間との関係

+ : BP法 □ : 関数誤差法1 × : 関数誤差法2 ◇ : 軌道誤差法

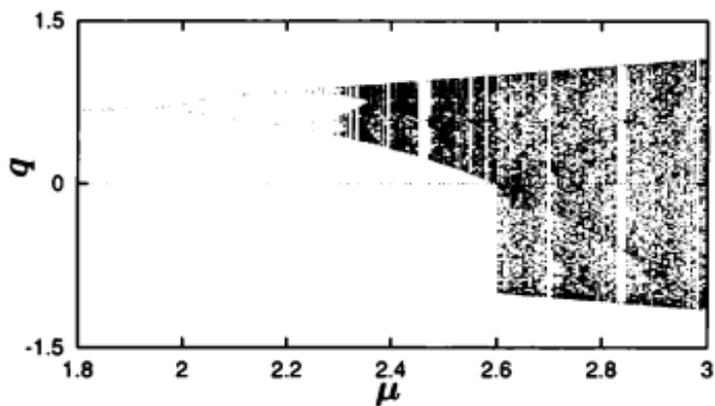


図 4-9 教師信号 $q(k+1) = \mu q(k)(1 - q(k)^2)$ の分岐図

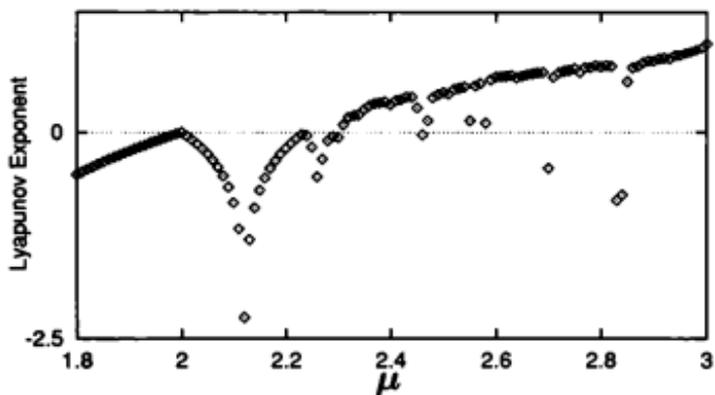


図 4-10 教師信号 $q(k+1) = \mu q(k)(1 - q(k)^2)$ のリアプノフ指数

あり、BP法と関数誤差法2の T_1 の値はほぼ等しい。また、関数誤差法1と軌道誤差法はともにBP法の約1/2の学習時間になっている。クライシス直前の $\mu = 2.55$ 付近から4方法ともに T_1 の値の変化が激しくなる。しかし、関数誤差法1の学習時間がBP法のそれの約1/2という関係と関数誤差法2とBP法の学習時間がほぼ等しいという関係は保たれており、 $T_2 = T_1$ とみなせば、式(4.3)の関係はパラメータ μ に依存せず、したがってアトラクタのサイズやリアプノフ指数にも依存せず、常に成立していることがわかる。

一方、軌道誤差法の場合はクライシス直前から T_1 が大きくなり、クライシスより後では μ の増加とともにBP法に近づいてゆき、BP法より大きくなる場合も生じている。これは、前述のようにリアプノフ指数が増加して教師アトラクタの軌道不安定性が大きくなるため、2ステップの軌道誤差を追跡することが困難になっていることを意味している。

次に、学習終了時間 T_2 を検討する。図4-8(b)を見ると、初期学習時間 T_1 とは異なり、アトラクタが周期点の場合でも分岐とともに T_2 は大きく変化することがわかる。

まず、 $\mu < 2.0$ でアトラクタが1周期点の場合は四つの方法とも T_2 はほぼ一定であり、関数誤差法、軌道誤差法ではともにBP法の約1/2になっている。また、BP法と関数誤差法2の T_2 はほぼ等しい。

μ が2.0を超えて2周期点の領域に入ると、 T_2 は1,000のオーダーから10,000のオーダーへと急激に増加する。これは、アトラクタ上で定義された関数を近似しているという立場で考えると、定義域が1点の関数と2点の関数との複雑さの違いの現れと解釈できる。アトラクタが2周期点の領域では、 μ の増加とともに T_2 は減少している。また、この領域では関数誤差法1の T_2 がBP法の約1/2という関係およびBP法と関数誤差法2の T_2 が等しいという関係は保たれたままであるが、軌道誤差法の T_2 は関数誤差法1のそれより大きくなっている。

μ をさらに増加させ、 $\mu = 2.31$ 付近でカオスの領域に入ると、 T_2 は再び急激に増加し、その後カオスの領域では μ の増加とともに T_2 は減少する。この領域でも、関数誤差法とBP法との関係は保たれているが、軌道誤差法の T_2 は関数誤差法1よりも、パラメータの値によってはBP法よりも大きくなっている。

$\mu = 3\sqrt{3}/2$ のクライシスにともない、BP法と関数誤差法1、2の T_2 は増加するが、軌道誤差法のそれは大きな変化を見せない。クライシスの後の領域でもBP法と関数誤差法1の T_2 は、2:1の関係を保ったまま μ の増加とともに緩やかに減

少する。また、この領域ではBP法と関数誤差法2に違いが生じ、BP法の T_2 が小さくなる。そして、この領域の大部分では、各方法の T_2 の大きさは次のような関係になっている。

軌道誤差法 < 関数誤差法1 < BP法 < 関数誤差法2

ただし、 μ の増加にともなう軌道誤差法の T_2 の減少率はBP法や関数誤差法のそれらより小さく、 μ が3.0に近づくと関数誤差法1と大差なくなっている。これは、 μ が大きくなるとアトラクタのリアプノフ指数が大きくなるので、カオスの効果が学習の加速よりも不安定化の方向により大きく働いたためと考えられる。

軌道誤差法の場合、クライシスの前後で T_2 はあまり変化しないが、学習曲線自体が変化しないわけではない。これを見るため、クライシス前後($\mu=2.59$ と $\mu=2.60$)の学習曲線の様子を図4-11に示す。この図より、四つの方法ともクライシスによって学習曲線は大きく変化し、BP法と関数誤差法では、学習が遅くなることわかる。軌道誤差法についても学習曲線は大きく変化し、学習初期の段階では関数誤差法と大差ないが、前述の学習の加速が生じるため T_2 は大きくならないのである。

以上、四つの方法による学習の終了時間 T_2 に関して、次のことが明らかとなった。

- (1) 四つの方法とも、分岐によるアトラクタの複雑化にともなって T_2 は大幅に増加する。すなわち学習に要する時間が長くなる。ただし、軌道誤差法については、クライシスの前後で学習曲線は大きく変化するが、学習の加速が生じるため T_2 は大きな変化を受けない。
- (2) アトラクタが2周期以上の周期点およびカオスの場合、パラメータ μ を増加させると、分岐が生じてアトラクタが定性的に変化するまでは T_2 は緩やかに減少する。
- (3) BP法と関数誤差法1の T_2 は、すべてのパラメータ値でほぼ2:1の関係を保っている。しかし、BP法と関数誤差法2の T_2 は、クライシス前のカオスの学習では、ほぼ等しいが、クライシスの後ではBP法や関数誤差法1による学習にも加速が生じ、関数誤差法2よりも学習が速く完了する。
- (4) 軌道誤差法の T_2 は、1周期点の場合は関数誤差法1と同じ、2周期以上の周期点とクライシス前のカオスの場合は関数誤差法1より大きく、クライシス後のカオスについては学習の加速のため関数誤差法1より小さくなっている。

分岐にともなう学習時間の増加については、学習している関数が複雑になるために生じるということで自然に解釈できる。すなわち、教師信号はアトラクタ上

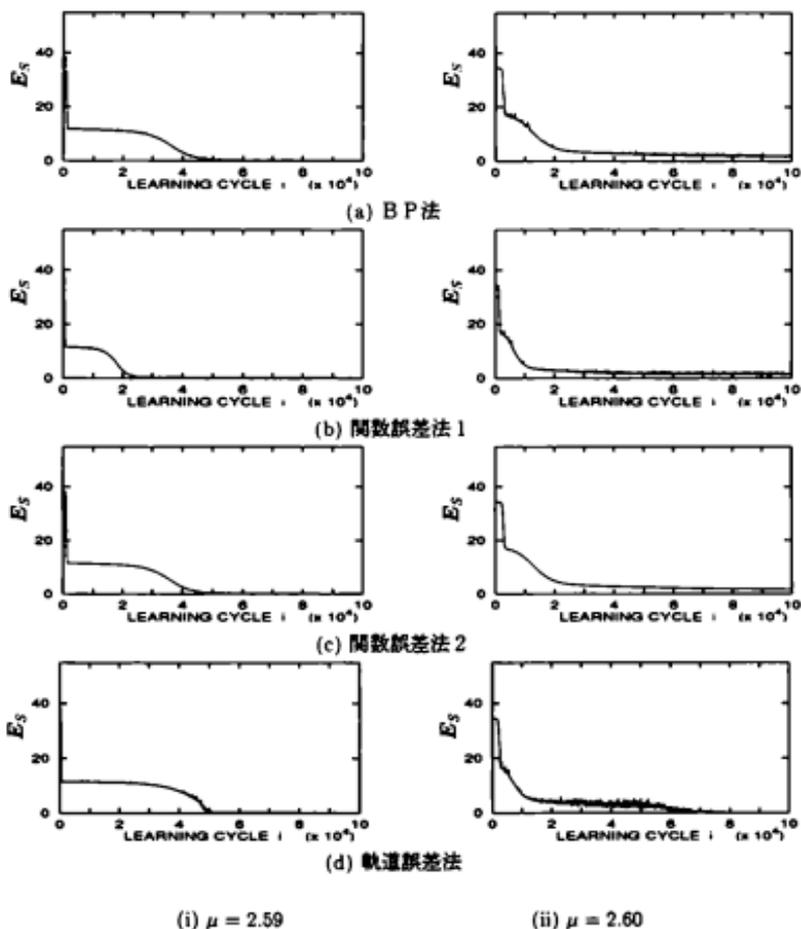


図 4-11 クライシス前後の学習曲線

でのみ与えられるので、定義域がアトラクタ上に制限された関数を学習していることに実質はなっている。したがって、分岐によりアトラクタが不動点、周期点、カオスと複雑化するにしたがって、学習している関数も複雑化するのである。

また、アトラクタが定性的に変化するまでは T_2 は μ の増加にともない減少するという現象については、関数の定義域の直径が大きくなるので、より広い範囲の情報が学習に利用できるようになるためと考えられる。

クライシスのあと軌道誤差法による学習時間が短くなるのは、 $\mu = 2.80$ の場合に説明した「カオスによる学習の加速」が生じているためである。また、実質的な学習係数 $N_B \cdot \eta$ の等しい B P 法と関数誤差法 2 に違いが生じているのは、B P 法による学習についても軌道誤差法と類似した加速が生じているためと考えられる。

以上の考察より、軌道誤差法における学習の加速現象は、クライシス後の原点对称なカオスが存在するパラメータ領域でのみ生じていることが判明した。また、この領域では B P 法や関数誤差法 1 による学習においても加速が生じており、B P 法が関数誤差法 2 より速くなっている。このことから、アトラクタの原点对称性と学習における揺らぎが学習の加速に関与しているものと推察される。

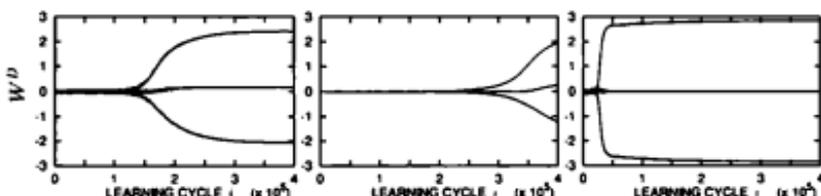
学習の加速とアトラクタの対称性との関連を調べるため、前述の B P 法、関数誤差法 2、軌道誤差法によるカオス ($\mu = 2.80$) の学習における学習曲線とシグモイド関数のバイアスパラメータ W_m^D の変化を測定し、図 4-12 に示した。

まず、関数誤差法 2 による学習での W_m^D の変化を見ると、 W_m^D は、零付近の微小かつランダムに与えた初期値の近傍に長く留まろうとする傾向にある。ところが、学習回数が 30 万の付近から二つの W_m^D が大きく変化し始め、これにともなって、誤差が減少することが併せて示した学習曲線よりわかる。この傾向は、B P 法と軌道誤差法でも見られ、B P 法では 20 万回、軌道誤差法では 2 万回付近で W_m^D が大きく変化し、学習が加速されている。すなわち、学習の加速には W_m^D の変化が大きく関与し、その変化する時期が学習の速度を決定している。

W_m^D の変化が学習速度を律する理由は次のように説明される。カオスアトラクタは原点对称であり、近似すべき関数は奇関数である。シグモイド関数も奇関数であり、RNN の出力関数もその加重和であるから、やはり奇関数である。したがって、シグモイド関数のバイアスである W_m^D が零でなくなることは、奇関数の近似にとっては都合が悪いように思われる。しかし、今、SU は 3 個であるから、すべてのシグモイド関数が奇関数である必要はなく、その加重和としての出力が奇関数であればよい。図 4-13 は関数誤差法 2 および軌道誤差法による 20 万回の学習後の R



(i) 学習曲線



(ii) W^D の変化 (上より、 W_1^D 、 W_3^D 、 W_2^D)

(a) BP法

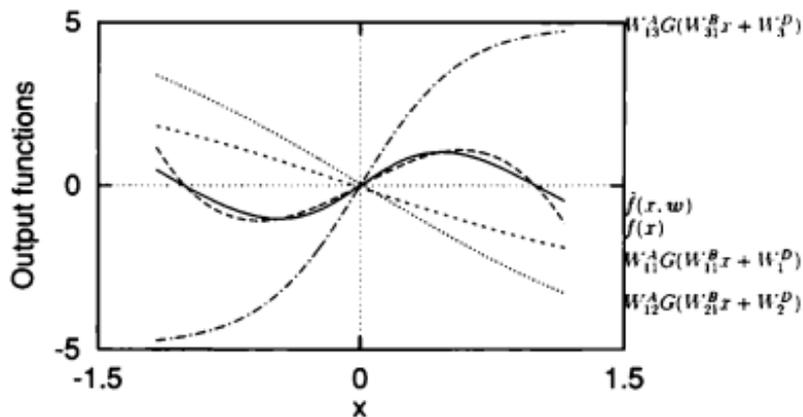
(b) 関数誤差法2

(c) 軌道誤差法

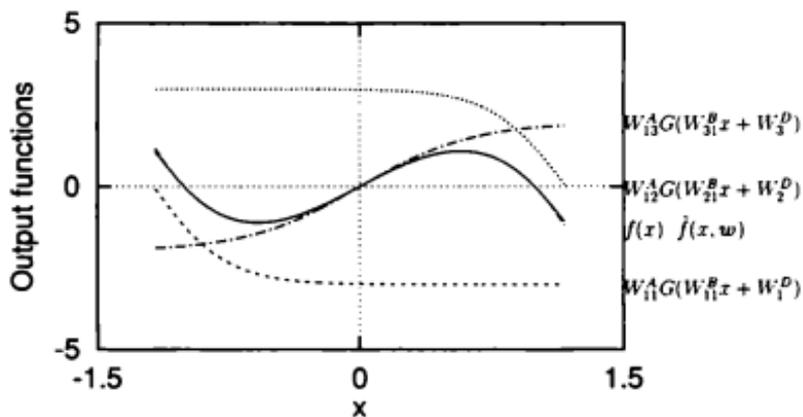
図4-12 学習曲線と W^D の変化

NNの出力関数と個々のシグモイド関数のグラフを表わしており、これを見ると、 W_m^D が零付近に留まり、シグモイド関数が三つとも奇関数である関数誤差法2による関数近似よりも、二つのシグモイド関数が原点对称性を失っているが、全体として対称性を保っている軌道誤差法による近似の方が優れていることが見てとれる。二つのシグモイド関数が原点对称性を失うことによりRNN全体としての関数合成の自由度が向上し、より高精度の近似が実現するのである。

それでは、三つの方法で W_m^D が変化し始める時期に違いが生じ、学習速度が異なるのはなぜであろうか。ここに、(逐次型)BP法と軌道誤差法の有する揺らぎの性質が関与するのである。



(a) 関数誤差法 2



(b) 軌道誤差法

図 4-13 RNN の出力関数

まず、関数誤差法による W_m^D の更新則を書き下すと、次のようになる。

$$\Delta W_m^D = -\eta \sum_{k=0}^{N_B-1} [\hat{f}(q(k), \omega) - f(q(k))] \times W_{1m}^A G'(W_{m1}^B q(k) + W_m^D). \quad (4.4)$$

W_m^D の初期値は微小かつランダムであるから、近似的に $W_m^D = 0$ と考える。したがって、 $G'(W_{m1}^B q + W_m^D)$ は q に関する偶関数である（なぜならば、 $G'(x) = 4/(e^x + e^{-x})^2$ は偶関数）。また、カオスアトラクタは原点对称であるから、その不変密度関数も偶関数である。さらに、RNN は奇関数 f を近似しているのだから、 $\hat{f}(q, \omega)$ も奇関数であると仮定できる。したがって、修正量 ΔW_m^D は、式(4.4)のように、正負均等に与えられるデータ点での奇関数の値の総和であるから、その1データ当りの値 (N_B で割った値) は N_B が大きいとほぼ零になる。したがって、 N_B が100という大きな値を持つ関数誤差法2では、 W_m^D が零付近の初期状態から移動しにくく、上に述べた理由により関数近似の精度が向上しないため、学習が遅れるのである。

次にBP法による学習を考える。BP法による学習での1サイクル当たりの W_m^D の変化は次式で与えられる。

$$\Delta W_m^D = -\eta [\hat{f}(q(k), \omega) - f(q(k))] \times W_{1m}^A G'(W_{m1}^B q(k) + W_m^D). \quad (4.5)$$

ここで、もし式(4.5)を教師信号の不変測度に関して平均すれば、それは、エルゴード性より関数誤差法の場合の W_m^D の変化(式(4.4))を近似することになる。これが、第3章で述べた確率的降下法の理論に基づく解析の帰結である。

しかし、平均的な学習の方向は関数誤差法2と一致しても、毎回の学習にはこの方向からの偏差が必ず生じ、これが揺らぎとなって W_m^D を零付近から移動させる働きをするものと考えられる。

軌道誤差法による学習では、BP法における逐次型学習の揺らぎに学習されたカオスの揺らぎが加わり、学習を一層加速している。軌道誤差法の場合の ΔW_m^D は次のようになる。

$$\Delta W_m^D = -\eta \frac{1}{2} \frac{\partial}{\partial W_m^D} \|\hat{f}(q(k), \omega) - f(q(k))\|^2 - \eta \frac{1}{2} \frac{\partial}{\partial W_m^D} \|\hat{f}^2(q(k), \omega) - f^2(q(k))\|^2 \quad (4.6)$$

式(4.6)の右辺第1項を Δ_1 、第2項を Δ_2 とすると、

$$\begin{aligned} \Delta_1 &= -\eta [\hat{f}(q(k), \omega) - f(q(k))] \times W_{1m}^A G'(W_{m1}^B q(k) + W_m^D) \\ \Delta_2 &= -\eta [\hat{f}^2(q(k), \omega) - f^2(q(k))] \times W_{1m}^A G'(W_{m1}^B \hat{f}(q(k), \omega) + W_m^D) \\ &\quad - \eta [\hat{f}^2(q(k), \omega) - f^2(q(k))] \times W_{1m}^A G'(W_{m1}^B q(k) + W_m^D) \\ &\quad \times \sum_{i=1}^M W_{i1}^A G'(W_{i1}^B \hat{f}(q(k), \omega) + W_i^D) W_{i1}^B. \end{aligned}$$

となる。 Δ_1 はBP法の ΔW_m^D (式(4.5)) に他ならない。つまり、この項は逐次型学習に起因する揺らぎを表わしている。そして、 Δ_2 が学習されたアトラクタのダイナミクスに起因する項である。実際、 Δ_2 の第2項は j の変分を表わしており、RNNがカオスのダイナミクスを獲得するにつれて、その軌道不安定性より、この変分の大きさは1のオーダーになる。

軌道誤差法による学習にともなう $|\Delta_1|$ と $|\Delta_2|$ の変化と学習曲線を図4-14に示す。この図より、学習曲線が振動し、学習が加速している付近で $|\Delta_2|$ が $|\Delta_1|$ の5~10倍の値になっていることがわかる。これが軌道誤差法による学習がBP法よりもさらに速くなる理由である。

4.2.6 結合荷重の収束過程

前項では、BP法と軌道誤差法による学習の加速がパラメータ W^D の変化に起因することを明らかにした。本項では、 W^D も含めたすべての結合荷重の学習に伴う変化過程を追跡し、学習法によるパラメータの収束過程の相違を検討する。

前項と同様、 $\mu = 2.80$ のカオスを、BP法、関数誤差法1 ($N_B = 2$)、関数誤差法2 ($N_B = 100$)、軌道誤差法 ($N_B = 2$) と確率的BP法 ($\eta = 0.01, \beta = 0.005$) の五とりの方法で400,000回学習させた場合の、結合荷重の収束過程を求め、図4-15に示した。

まず、BP法と関数誤差法1を比較すると、両者の結合荷重はすべて速度比1:2の関係で変化し、ほぼ同じ値に収束することがわかる。したがって、確率的降下法による学習速度の評価は、誤差の減少速度だけでなく、個々のパラメータの収束速度に関しても成立していることがわかる。

BP法と関数誤差法2を比較しても、学習の初期段階では両者の間に相違はほとんど見られない。前述のように、 W^D の対称性が崩れ、学習が加速される時期が関数誤差法2の場合にBP法より遅れる。このため、400,000回の学習では関数誤差法2による学習は最終段階に到達しておらず、結合荷重空間での最終的な収束先はBP法の場合にはほぼ一致するものと考えられる。

軌道誤差法については、学習初期段階では三つの結合荷重ともBP法とはほぼ同じ変化をしているが、学習が加速する $i = 20000$ 付近(図4-7(b)参照)からBP法とは異なった収束過程をたどり、ほぼ最終的とみなせる400,000回学習後の状態もBP法とは異なっている。また、確率的BP法の場合も結合荷重の収束先はBP法や軌

* 結合荷重の収束先を比較する際、中間層のノードを表わす添字 m の順序を入れ替えたもの

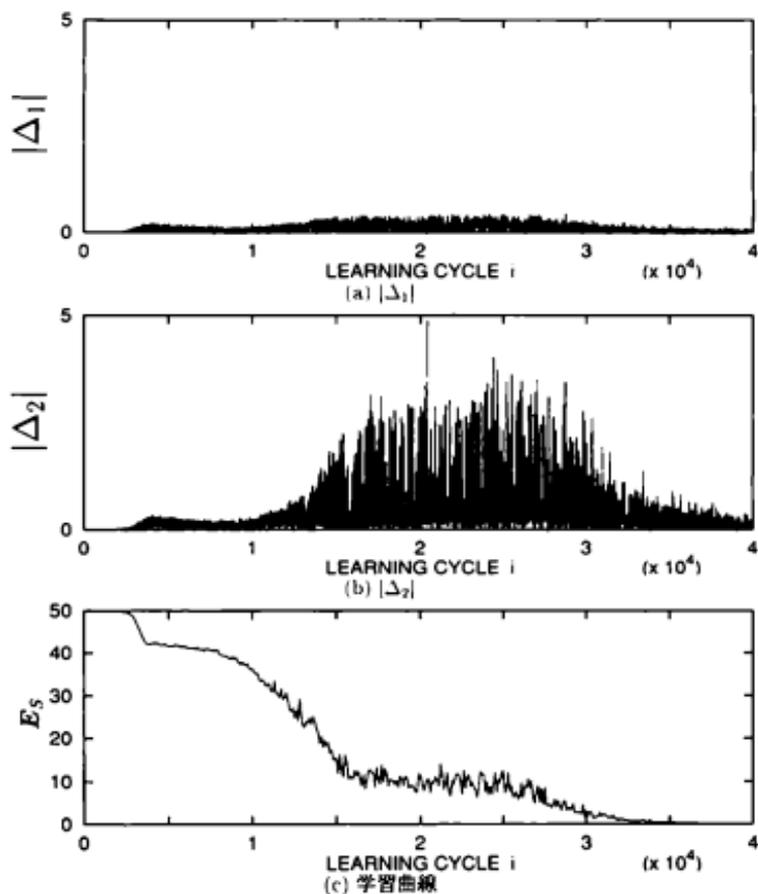


図 4-14 軌道誤差法による学習における $|\Delta_1|$ と $|\Delta_2|$ の変化と学習曲線

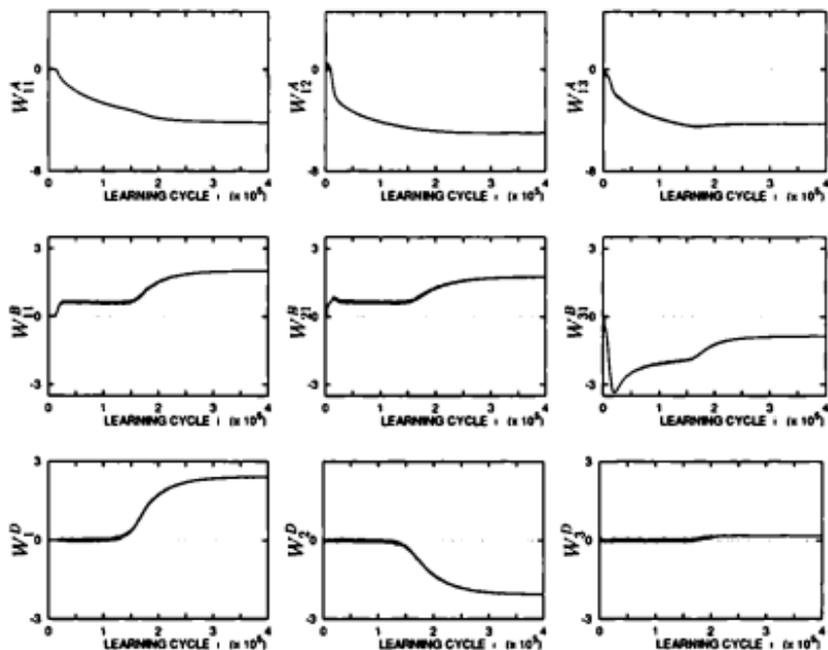


図 4-15 結合荷重の変化 (a) BP法

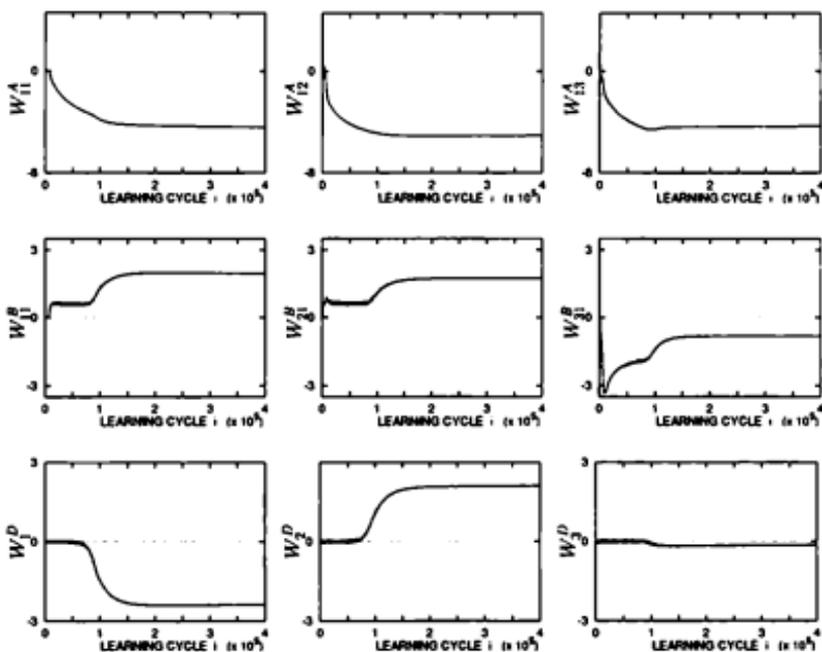


図 4-15 結合荷重の変化 (続き) (b) 関数誤差法 1

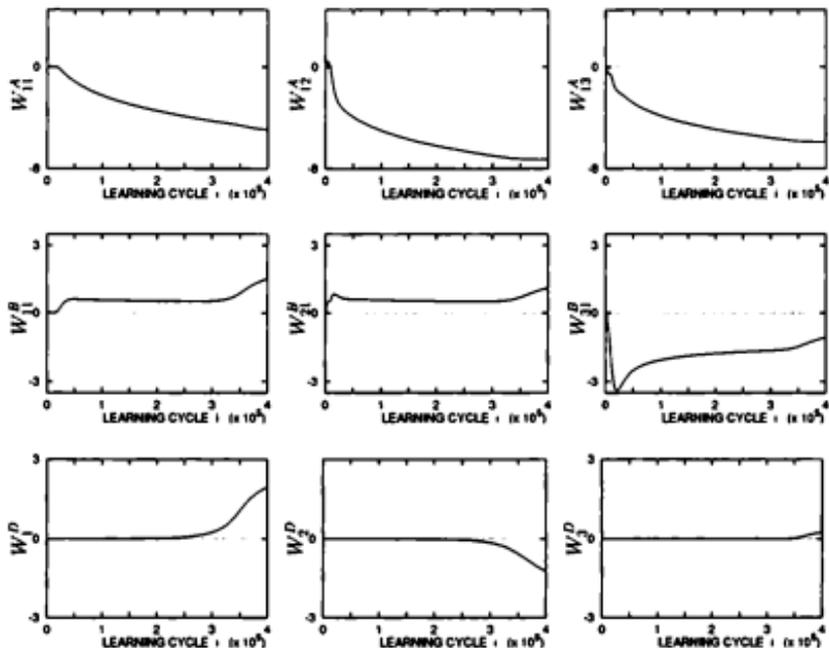


図 4-15 結合荷重の変化(続き) (c) 関数誤差法 2

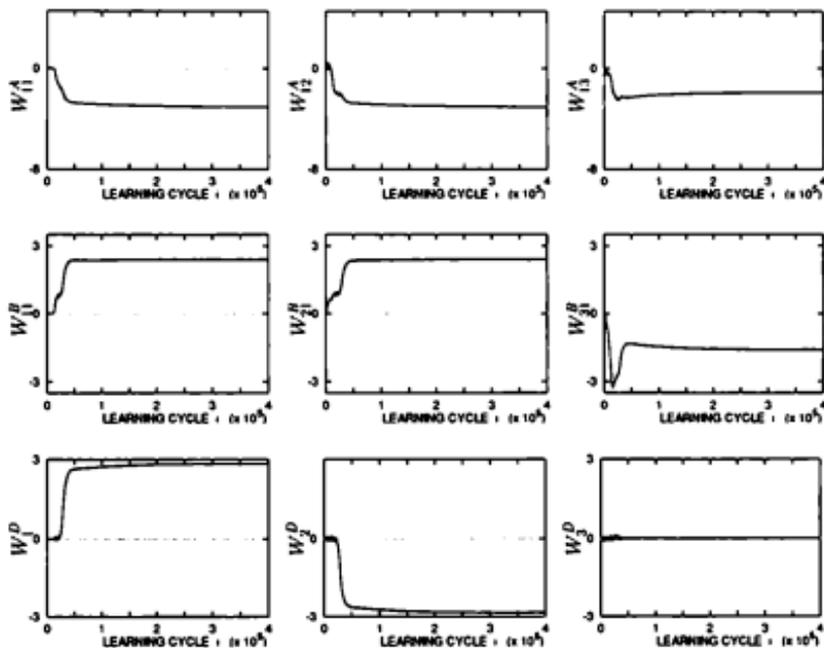


図 4-15 結合荷重の変化(続き) (d) 軌道誤差法

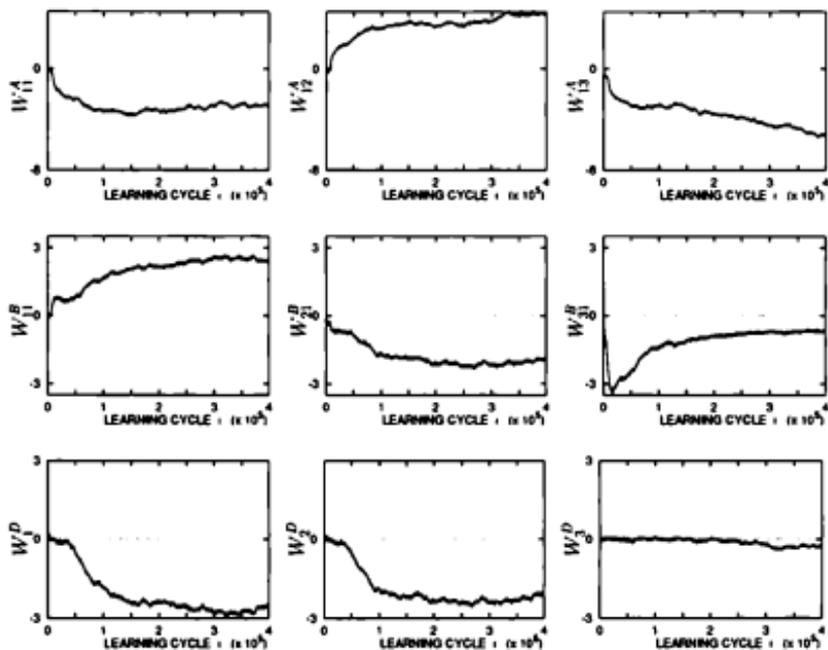


図 4-15 結合荷重の変化(続き) (e) 確率的BP法

道誤差法のそれらとは異なっている。

以上より、軌道誤差法と確率的BP法は、結合荷重の空間でともにBP法とは異なった状態に収束することがわかった。このことから、軌道誤差法や確率的BP法による学習では、BP法や関数誤差法の陥るローカルミニマムを回避していると考えられることもできる。実際、400,000回学習後の誤差評価関数の値はBP法、関数誤差法1、関数誤差法2、軌道誤差法、確率的BP法の場合、それぞれ、約0.010、0.009、0.174、0.002、0.120であり、軌道誤差法のものが最も小さくなっている。しかし、結合荷重空間は多くのStiefel's Cage³⁶⁾状の谷が交錯した極めて複雑な構造を有しているものと考えられるので³¹⁾、ローカルミニマムとグローバルミニマムの相違だけでなく、極小点に接近しているか否かを判定することさえ数値的には困難である。

結合荷重の収束過程と学習曲線から確実に判断できるのは、以下のことである。すなわち、学習過程を結合荷重空間上の力学系と考えると、すべての W_{ij}^D が0である状態が一つの不变集合になっており、学習の初期段階では、どの学習方法を用いてもこの不变集合内で学習が進行する。この不变集合はそれ自体一つのStiefel's Cageになっており、非常に緩やかに傾斜した底をもつ谷になっているであろう。したがって、この谷底のどこかにグローバルミニマムが存在する可能性も否定できない。しかし、前述のように、この集合内では誤差を一定以上減少させることが困難になっている。関数誤差法2は近似的な最急降下法であるので揺らぎが小さく、この不变集合が比較的安定になっているため、ここからなかなか脱出できず、学習が遅れる。しかし、軌道誤差法や確率的BP法では揺らぎが大きく、容易に脱出できるのである。学習における揺らぎは、この不变集合を不安定化し、そこからの脱出の時期を早め、学習を促進する働きをしている。揺らぎの大きさは、脱出速度を左右するが、脱出後の収束先を保証するものではない。したがって、早期に学習が進行する確率的BP法の最終的な誤差が、上記のようにBP法や関数誤差法1のそれらよりも大きくなることも起こり得るのである。

4.3 教師信号の提示方法について

時系列の学習はその時系列を生成する力学系(関数)の同定であるという観点からは、学習において、時系列を意識した特別なアルゴリズム上の工夫は必要なのは一視すべきであるが、順序をどのように入れ替えても軌道誤差法の結果はBP法とは一致しないので学習の最終状態は異なっていると判断できる。

いように思われる。実際、確率的降下法の理論に基いた考察により、時系列学習アルゴリズムは一般に、平均的な意味ではすべて逐次型のBP法に等価であることを本論文の第3章で既に示した。

しかし、前節では、原点対称な1次元関数から生成されるカオスを学習する場合、逐次型BP法の持つ揺らぎや、学習されたカオスの不安定性が学習を加速することを見出した。また、本堂ら³²⁾はカオスの学習においては、平均的に扱うと見失われる、逐次型のBP法の揺らぎとカオスアトラクタに含まれる不安定周期点が学習の促進に重要な役割を演じていることを発見し、一様分布白色雑音とカオス時系列の相違を論じている。このように、学習されるべき力学系は静的な関数にすぎないとしても、その学習過程においては時系列の種類、教師信号としての提示方法、学習アルゴリズムなど動的な性質がさまざまな影響を持つものと考えられる。

ここでは、前節に引き続き、三層階層型のRNNによる1次元離散力学系の学習を対象として、教師信号時系列のデータ数やその分布密度、時間相関が学習過程におよぼす影響を考察する。

学習対象は、Cubic map、

$$q(k+1) = 2.8q(k)(1 - q(k)^2) \quad (q \in \mathbb{R}), \quad (4.7)$$

および Logistic map、

$$q(k+1) = 4.0q(k)(1 - q(k)) \quad (q \in \mathbb{R}) \quad (4.8)$$

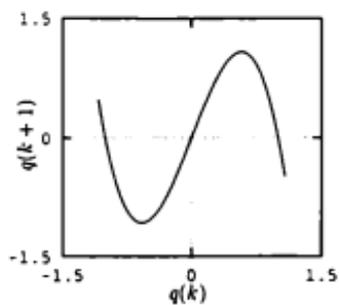
で記述される1次元系とする。これらは図4-16、17にそれぞれ示すカオスアトラクタおよび不変密度分布を有しており、このカオス時系列を教師信号として次節以下の学習実験を行なう。以下、この二つの写像を教師関数と呼ぶ。

4.3.1 有限時間時系列データを用いた学習

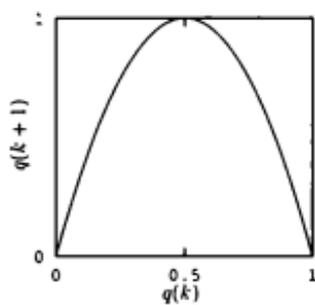
学習方法

前節の二つの教師関数の各々について、その反復によって N 個の時系列データ $\{q(0), q(1), \dots, q(N-1)\}$ を作成しておき、これを教師信号として、

- (a) BP法 ($N_B = 1$)
- (b) 関数誤差法1 ($N_B = N - 1$)
- (c) 関数誤差法2 ($N_B = 2$)

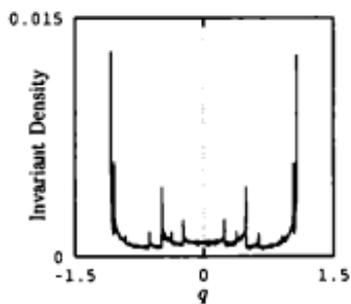


(a) Cubic map

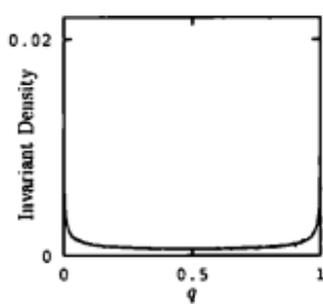


(b) Logistic map

図 4-16 教師関数



(a) Cubic map



(b) Logistic map

図 4-17 教師関数の不変密度分布

(d) 軌道誤差法 ($N_B = 2$)

の四とおりの方法で学習する。RNNの中間層ノード数は $M = 3$ とする。学習アルゴリズムは第2章2.3節で述べた、[Step1]~[Step3]から成るアルゴリズムを使用する。ただし、[Step3]において、 $k = N_B$ ではなく、 $k = 1$ を時刻0にリセットすることに変更する。そして、[Step1]~[Step3]のサイクルを $(N - N_B)$ 回繰り返したあと、 $k = 1$ のデータは最初の $q(0)$ からもう一度使用し、以下これを反復する。たとえば、 $N = 5$ 、 $N_B = 2$ の場合は各サイクルで使用するデータは、 $\{q(0), q(1), q(2)\}$ 、 $\{q(1), q(2), q(3)\}$ 、 $\{q(2), q(3), q(4)\}$ 、 $\{q(3), q(4), q(5)\}$ 、 $\{q(0), q(1), q(2)\}$ 、 $\{q(1), q(2), q(3)\}$... のように変化していく。基本的には $N = 5$ 個のデータを循環させて使用するのであるが、軌道誤差法では時系列の連続性がなければ意味がないので、 $\{q(4), q(5), q(0)\}$ のようなデータは使用せず、 $\{q(3), q(4), q(5)\}$ の次は $\{q(0), q(1), q(2)\}$ に戻るのである。また、このとき、(b) の関数誤差法1は、 $N - 1$ 個の点、 $\{q(0), \dots, q(N - 2)\}$ での関数 f を近似する一括型のBP法になる。また、学習係数 η は、(a)~(d)のすべての方法について、 $\eta = 0.02/N_B$ とする。 $N_B \cdot \eta$ が一定なので、(a)~(c)はすべて平均的には等価になる。

学習の評価基準

学習の進行を評価する量として、式(4.2)で定義される、アトラクタ上の二乗関数誤差 E_S を用いる。 E_S の減少により学習の進行状況が把握できる。ここでは、 $N_S = 200$ とし、また、結果は E_S を N_S で割った平均誤差で表示する。

次に、(i) 学習の速度、(ii) 学習後のRNNの時系列予測能力、(iii) アトラクタの統計的性質の獲得能力をそれぞれ評価する指標として、(i) 学習終了時間、(ii) 相対二乗平均誤差²¹⁾、(iii) リアプノフ指数を以下のように定義する。

(i) 学習終了時間 T :

E_S が学習開始時点の1%になる学習サイクル数を学習終了時間 T と定義する。

(ii) 相対二乗平均誤差 $RMSE(k)$:

$$RMSE(k) = \frac{MSE(k)}{\sigma} \quad (k = 0, 1, 2, \dots), \quad (4.9)$$

$$MSE(k) = \sqrt{\frac{1}{L} \sum_{l=1}^L (x(l+k) - q(l+k))^2},$$

$$\sigma = \sqrt{\frac{1}{L} \sum_{l=1}^L (q(l) - \bar{q})^2}, \quad \bar{q} = \frac{1}{L} \sum_{l=1}^L q(l)$$

とする。ただし、 x は学習後のRNNの出力であり、各 l について $x(l) = q(l)$ のように初期値を教師信号 q に一致させ、 $x(l+k)$ ($k = 0, 1, 2, \dots$) を計算していくものとす

る。つまり、 $RMSE(k)$ は同一初期値から出発した RNN の軌道と教師信号との k ステップ後の二乗誤差を L 個の初期値に関して平均したものである。本実験では、 $L = 10,000$ とした。

(iii) リアプノフ指数 λ :

次式で定義される学習終了後の RNN のリアプノフ指数を用いる。

$$\lambda = \lim_{k \rightarrow \infty} \frac{1}{k} \log \left\| \frac{d}{dq} f^k(q(0), \omega) \right\|.$$

実験結果と考察

まず、Cubic map の学習実験結果について述べる。

データ数 N に対する学習終了時間 T 、3 ステップ後の相対二乗平均誤差 $RMSE(3)$ 、リアプノフ指数 λ の変化を、図 4-18~20 にそれぞれ示す。学習サイクル数の上限を 10 万回に設定し、この回数で誤差が 1% にならない場合は、 $T = 100,000$ と表示する。また、相対二乗平均誤差とリアプノフ指数は 10 万回学習後の RNN について計算した。時系列データ数は、 $N = 5, 6, 7, 8, 9, 10, 11, 21, 31, 41, 51, 61, 71, 81, 91, 101$ とし、データ数が無制限の場合（関数誤差法 1 では $N_B = 100$ とする）も併せて図示している。

学習終了時間 T は、 $N < 11$ の場合は各方法とも上限値 10 万を示しており、この程度のデータ数では学習が十分進行しないことが確認できる。 $N = 11$ で T は急速に減少する。これより大きな N の値では、(d) の軌道誤差法が他の 3 方法に比べて学習が速く終了していることがわかる。また、(a)~(c) の 3 方法については、 $N \geq 21$ の範囲では N とともに T が増加していく。軌道誤差法についてはこのような傾向はみられず、 $N \rightarrow \infty$ になるにしたがって、

(d) 軌道誤差法 > (a) B P 法 > (c) 関数誤差法 2 > (b) 関数誤差法 1

の順に学習速度が小さくなる。これは、前節で述べた、教師信号のデータが無制限に与えられる場合の学習の加速現象が、データ数 N が 20 程度ですでに観測されることを意味している。 N の増加とともにデータの正負均等性が強まるため、B P 法と関数誤差法では W_m^D に加わる揺らぎが小さくなり、学習速度が減少することを図 4-18 は示している。また、このことは、図 4-21 に示した軌道誤差法の学習曲線からも確認できる。データ数 N が 11 になると軌道誤差法に特有のゆらぎがみられ、学習が急激に進行していることが見てとれる。

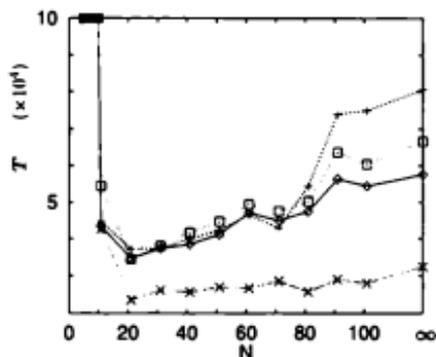


図 4-18 学習終了時間 (Cubic map)

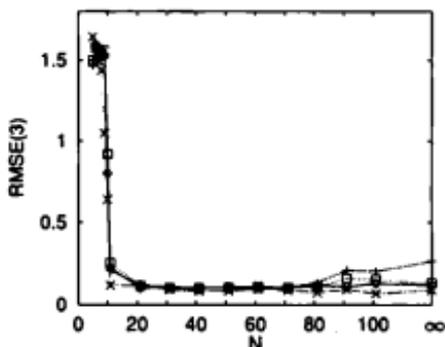


図 4-19 相対二乗平均誤差 (Cubic map)

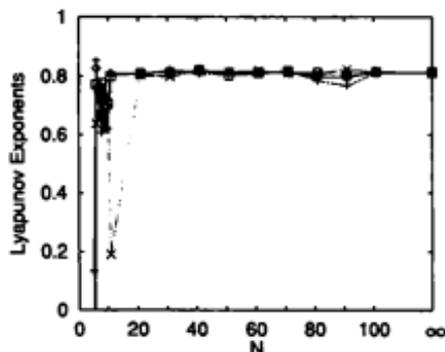
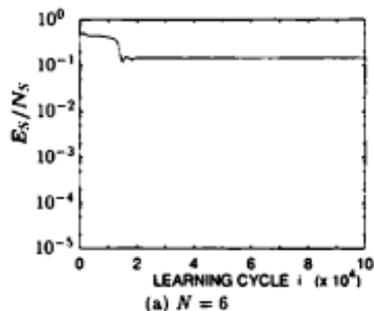
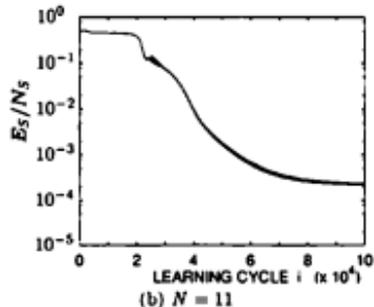


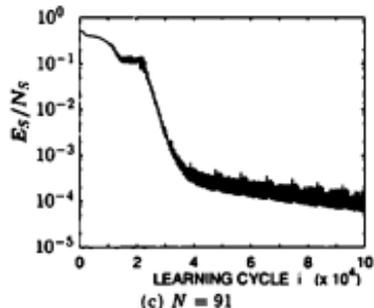
図 4-20 リアプノフ指数 (Cubic map)



(a) $N_B = 6$



(b) $N_B = 11$



(c) $N_B = 91$

図 4-21 学習曲線 (Cubic map)
軌道誤差法 ($N_B = 2$)

図 4-18 ~ 図 4-20 および 図 4-22 ~ 図 4-24 において、
 \circ : BP 法 ($N_B = 1$)
 $+$: 関数誤差法 1 ($N_B = N - 1$)
 \square : 関数誤差法 2 ($N_B = 2$)
 \times : 軌道誤差法 ($N_B = 2$) を変わす。

図4-19の相対二乗平均誤差をみると、 $N \geq 81$ の領域では各学習法の時系列予測能力は学習速度と同様、 $(d) > (a) > (c) > (b)$ の順に低くなっていることがわかる。一方、図4-20に示したようにリアブノフ指数は、 $N \geq 11$ では各方法とも大差なく、教師信号の値 $\lambda = 0.806$ をよく近似している。これらの結果は、時系列の予測を行なうためには高精度の学習が要求されるが、教師アトラクタの統計的性質を獲得するだけなら、比較的小数のデータで十分であることを示唆している。

次に、Logistic map の学習結果について述べる。

データ数 N に対する学習終了時間 T 、相対二乗平均誤差 $RMSE(3)$ 、リアブノフ指数 λ の変化を、図4-22~24にそれぞれ示す。学習サイクル数の上限を20万回に設定し、この回数で誤差が1%にならない場合は、 $T = 200,000$ と表示する。また、相対二乗平均誤差とリアブノフ指数は20万回学習後の RNN について計算した。時系列データ数はCubic map の場合と同じである。

学習終了時間 T は各方法とも $N = 10$ 付近から減少しているが、Cubic map の場合と異なり、軌道誤差法のみが大きな値となっている。他の3方法には大差なく、 N とともに増加するような傾向もみられない。これは、Logistic map のアトラクタはCubic map のような原点対称性を持たないため、上述のような逐次型BP法や軌道誤差法による揺らぎが学習に及ぼす影響が小さいためと考えられる。したがって、 $N_B \cdot \eta$ が一定のBP法と関数誤差法は確率的降下法の理論どおり同程度の速度で学習が進行する。 $N_B = 2$ の軌道誤差法では、2ステップめの誤差を評価する成分がCubic map の場合には W_m^D に揺らぎを与え学習を加速するのであるが、Logistic map の学習では単なるノイズの効果しか持たないため、学習の進行が遅れるのである。

一方、相対二乗平均誤差とリアブノフ指数については軌道誤差法による学習が最も良好な結果を得ている。これは、図4-25に示す $N = 101$ の場合の学習曲線からわかるように、誤差が1%まで減少する学習の初期段階では、軌道誤差法は他の方法に比べて進行が遅いが、学習の最終段階での誤差減少率が著しく大きく、20万回の学習後では、もっとも誤差が小さくなっているためである。この学習の最終段階での軌道誤差法による学習の加速はCubic map の学習の場合に観測されたものとは異質のものと思われる。

BP法、関数誤差法のリアブノフ指数が負値になっているのは、 RNN はLogistic map の関数形を十分獲得したが、このアトラクタはクライシスの瞬間のものであるため、その安定性の学習がうまくできなかったためである。

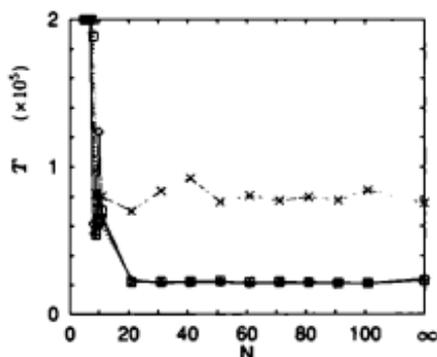


図 4-22 学習終了時間 (Logistic map)

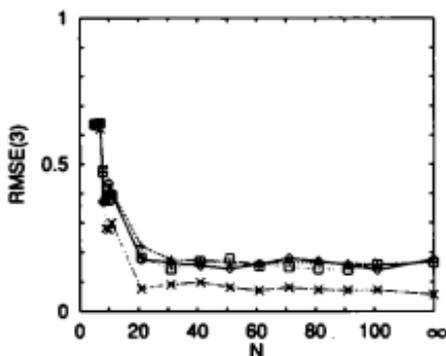


図 4-23 相対二乗平均誤差 (Logistic map)

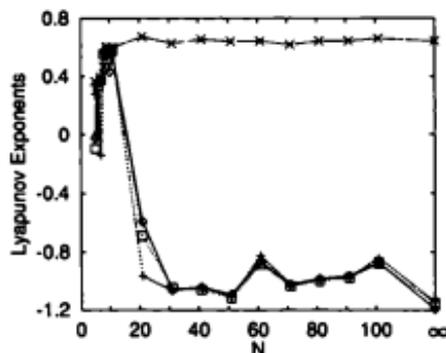
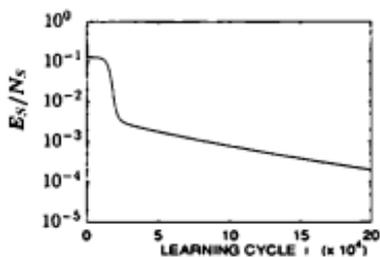
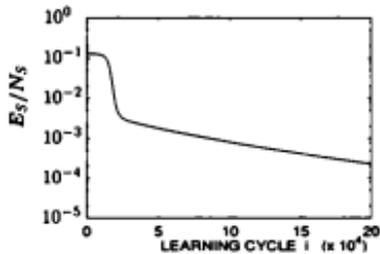


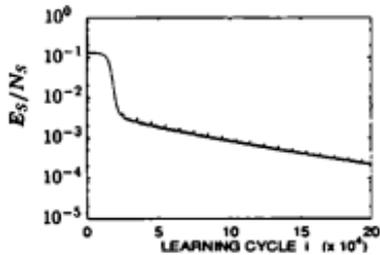
図 4-24 リアプノフ指数 (Logistic map)



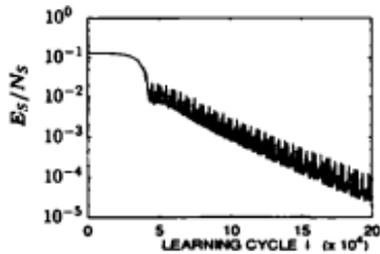
(a) BP 法 ($N_B = 1$)



(b) 関数誤差法 1 ($N_B = N - 1$)



(c) 関数誤差法 2 ($N_B = 2$)



(d) 軌道誤差法 ($N_B = 2$)

図 4-25 学習曲線 (Logistic map $N = 101$)

4.3.2 時系列の分布密度・時間相関を変化させた学習

学習方法

時系列 $\{q(k)\}$ を以下の四とおりの方法で作成し、教師パターンとして $(q(k), f(q(k)))$ を順次提示して、BP法により関数 f を学習する。学習係数は $\eta=0.02$ とする。

- (a) 基本時系列: 教師関数の反復により生成される時系列そのもの。
- (b) 不変密度分布白色雑音: (a) の基本時系列を30ステップおきにサンプリングして得られる時系列。これはカオスの混合性より、時間相関がほぼ消滅し、基本時系列と同一の不変密度分布にしたがう白色雑音になる。
- (c) 一様分布白色雑音: 基本時系列の作るカオスアトラクタの領域内で一様に分布する白色雑音。
- (d) 異なる関数から生成される時系列: Logistic map の生成する時系列をその最大値・最小値が Cubic map のアトラクタの最大値・最小値になるようにサイズを修正し、Cubic map の学習に用いる。Logistic map の学習の場合は逆に Cubic map の基本時系列を Logistic map のアトラクタのサイズに合わせて用いる。

実験結果と考察

Cubic map と Logistic map の上記4方法による学習曲線を図4-26、27にそれぞれ示す。Cubic map の学習では学習の速さは、

- (d) Logistic 時系列 > (a) 基本時系列 \approx (b) 不変密度白色雑音 > (c) 一様白色雑音

の順になる。

まず、(a) と (b) とに差がみられないことから、自己相関関数で測定できるような教師信号の時間相関は学習速度にあまり影響しないことがわかる。しかし、時系列の密度分布、すなわち教師パターン $(q, f(q))$ を与える点 q の出現頻度は大きく影響している。(c) の一様分布に比べ(a)、(b)の不変密度分布の方が速い。しかし、(d)の Logistic 時系列はさらに速くなっている。Logistic map の不変密度分布は図4-17(b)に示したようにアトラクタの両端で高くなっている。Cubic map の密度分布も両端で最大になるが、その他にもピークを持っている(図4-17(a))。このことから、アトラクタの両端を強調してデータを提示した方が学習の進行が速いと考えられる。シグモイド関数は単調増加であるから、その加重和で Cubic map を近似する場合、 f が減少する部分を強調して学習した方が効率が良いということである。

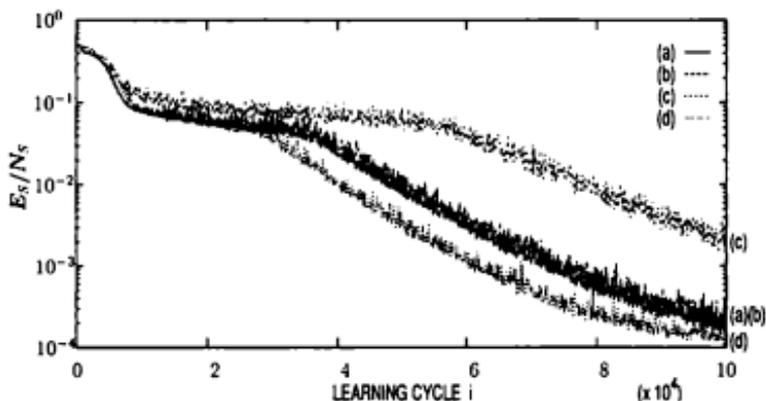


図 4-26 時系列の分布密度・時間相関を変化させた B P 法による学習曲線 (Cubic map)

- (a) Cubic map 時系列、(b) 不変密度分布白色雑音、
 (c) 一様分布白色雑音、(d) Logistic map 時系列

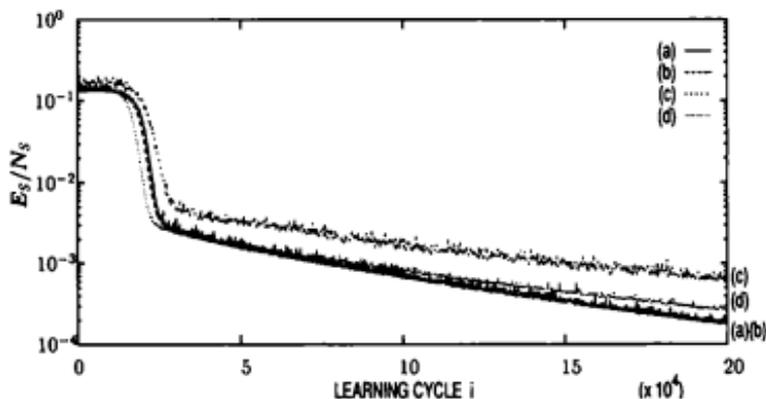


図 4-27 時系列の分布密度・時間相関を変化させた B P 法による学習曲線 (Logistic map)

- (a) Logistic map 時系列、(b) 不変密度分布白色雑音、
 (c) 一様分布白色雑音、(d) Cubic map 時系列

Logistic map の学習では学習の速さは、

(a) 基本時系列 \approx (b) 不変密度白色雑音 $>$ (d) Cubic 時系列 $>$ (c) 一様白色雑音

の順になる。やはり、時系列相関の影響はほとんど見られない。また、この場合も Logistic 時系列すなわち基本時系列を提示した学習が最も速い。やはりアトラクタの端点を重点的に学習した方が効率的であることがわかる。

4.4 連続力学系の学習実験

この節では、連続力学系の学習実験結果について述べる。学習対象は、2次元系の例として van der Pol 方程式、3次元系の例として Rössler 方程式をとりあげる。

以下の数値計算では、常微分方程式の求解には4次のルンゲクッタ法を使用している。ただし、波形誤差法におけるラグランジュ未定乗数の逆時間方向の求解(第2章式(2.22))にはオイラー法を使用している。

4.4.1 van der Pol 方程式の学習

van der Pol 方程式、

$$\begin{aligned}\frac{dq_1}{dt} &= q_2, \\ \frac{dq_2}{dt} &= \mu(1 - q_1^2)q_2 - q_1\end{aligned}\quad (4.10)$$

は、パラメータ $\mu = 1.0$ のとき、図4-28に示したリミットサイクルを持つ。以下では、このリミットサイクルを教師信号とした学習を考える。

速度誤差法による学習過程

速度誤差法を用いた学習で、学習係数 η と1サイクルの学習時間 T_B 、学習に要する時間 T_C の関係を調べる。 T_C は離散系の学習の場合と同様に定義する。すなわち、式(4.2)のように二乗関数誤差 E_S を定義し、 E_S が一定のしきい値 E_C を切る時間を T_C とする。

図4-28のリミットサイクルを速度誤差法で学習させた場合の学習曲線を図4-29に示す。図4-29の(a)、(b)、(c)は、それぞれ、

(a) $\eta = 0.01, T_B = 0.1$ 、(b) $\eta = 0.04, T_B = 0.1$ 、(c) $\eta = 0.01, T_B = 0.4$

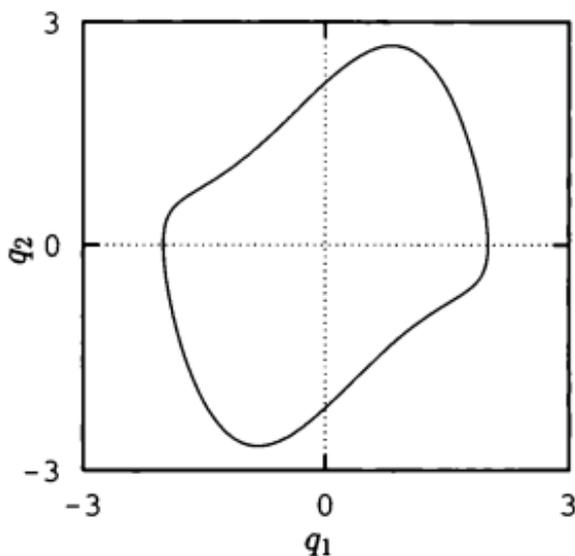
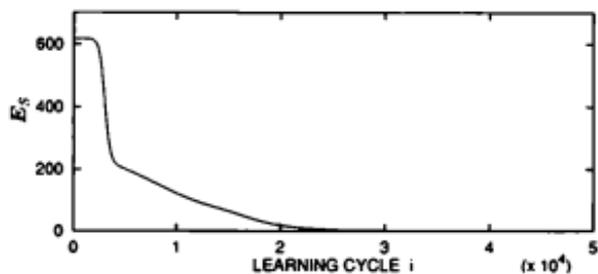
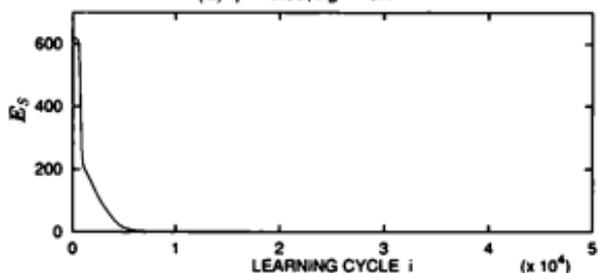


図4-28 van der Pol 方程式のリミットサイクル

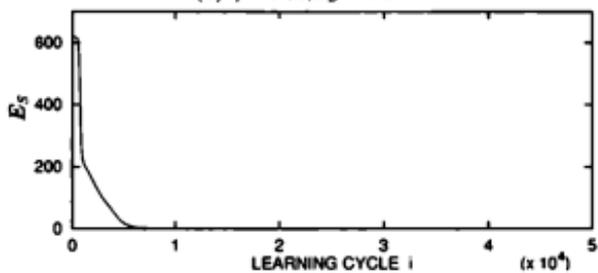
の三とおりの条件での学習曲線をそれぞれ示している。この結果を見ると、(b)と(c)の方が(a)に比べて学習が速く、(b)と(c)の学習曲線は学習初期の段階を見る限りほとんど一致していると言える。このことから、学習係数 η を4倍することと、学習時間 T_B を4倍することとは、ほぼ等価であることがわかる。この等価関係を明確にするため、学習初期段階における η と T_B および学習時間 T_C の関係を求め、図4-30に示した。図4-30(a)は T_B と T_C との関係を η をパラメータとして、(b)は η と T_C との関係を T_B をパラメータとして描いたものである。 E_C は50.0とした。



(a) $\eta = 0.01, T_B = 0.1$

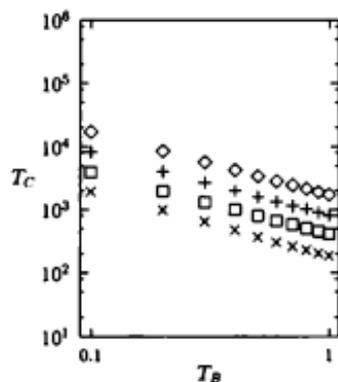


(b) $\eta = 0.04, T_B = 0.1$



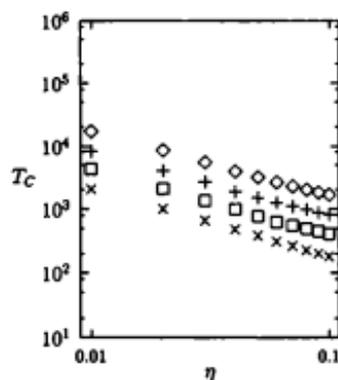
(c) $\eta = 0.01, T_B = 0.4$

図 4-29 速度誤差法による van der Pol 方程式の学習曲線



(a) T_B と T_C との関係

\diamond : $\eta = 0.01$, $+$: $\eta = 0.02$, \square : $\eta = 0.04$, \times : $\eta = 0.08$



(b) η と T_C との関係

\diamond : $T_B = 0.1$, $+$: $T_B = 0.2$, \square : $T_B = 0.4$, \times : $T_B = 0.8$

図 4-30 速度誤差法における学習パラメータと学習時間との関係

この図より、

$$T_C \propto \frac{1}{T_B \cdot \eta} \quad (4.11)$$

の関係が成立していることがわかる。これは、第3章で導出した理論式(3.33)に一致する。

波形誤差法による学習過程

次に波形誤差法を用いて同じく van der Pol 方程式のリミットサイクルを学習させた。設定条件は、

$$(a) \eta = 0.1, T_B = 0.1, (b) \eta = 0.8, T_B = 0.1, (c) \eta = 0.1, T_B = 0.2$$

の三とおりとし、その学習曲線を図4-31(a)、(b)、(c)にそれぞれ示した。

結果を見ると、(a)は(b)および(c)に比べて学習が速く、(b)と(c)はほぼ一致している。このことから、 η を8倍することと T_B を2倍することとは等価であることが推測でき、第3章で導出した関係、

$$T_C \propto \frac{1}{T_B^2 \cdot \eta} \quad (4.12)$$

の成立していることがわかる。この関係を確認するため、速度誤差法の場合と同じく、学習初期段階における η と T_B および学習時間 T_C の関係を求め、図4-32に示した。図4-32(a)は T_B と T_C との関係を η をパラメータとして、(b)は η と T_C との関係を T_B をパラメータとして描いたものである。しきい値 E_C は50.0である。この図から式(4.12)の成立していることが確認できる。

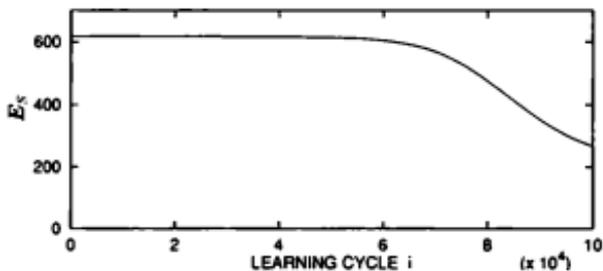
軌道追従能力の比較

学習が完了したRNNによる教師信号の軌道の追従能力を調べる。まず、同一初期値から出発した教師信号とRNNの軌道との誤差の時間変化を次のように定義する。

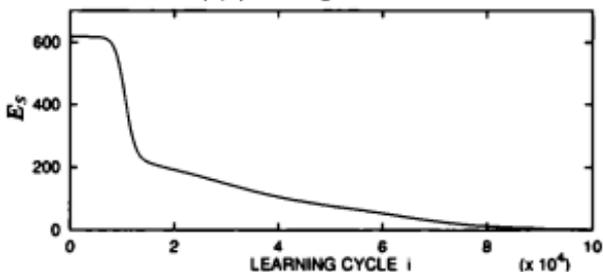
$$E_o(t) = \frac{1}{N_o} \sum_{j=1}^{N_o} \|q(x_j, t) - x(x_j, t, w)\|^2. \quad (4.13)$$

ここで、 x_j ($j = 1, \dots, N_o$) はアトラクタ上のサンプリング点である。第2章で述べたように、 $q(x_j, t)$ と $x(x_j, t, w)$ はそれぞれ、 x_j を初期値とする教師信号とRNNの軌道を表わす。

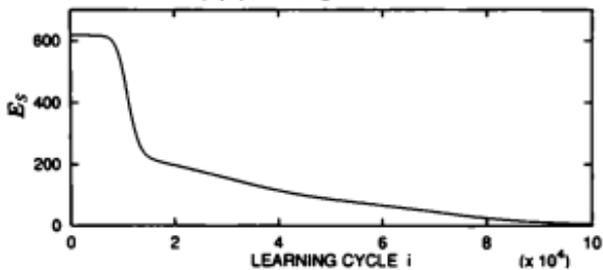
以下、速度誤差法および波形誤差法で1サイクルあたりの学習時間 T_B を変えた場合の $E_o(t)$ を比較する。初期値のサンプル数は $N_o = 100$ とし、また二乗関数誤差が $E_S = 1.08$ となった時点を学習完了とみなす。



(a) $\eta = 0.1, T_B = 0.1$

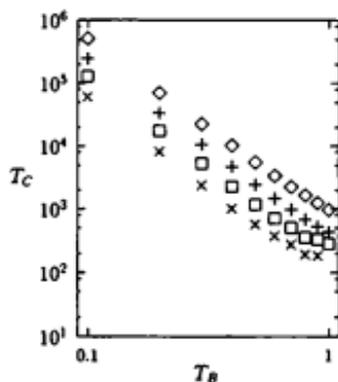


(b) $\eta = 0.8, T_B = 0.1$



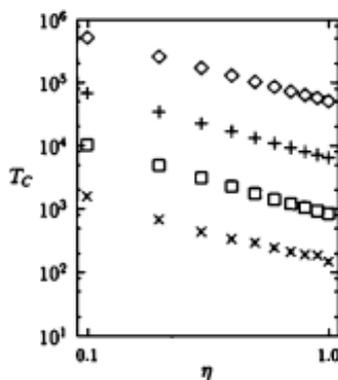
(c) $\eta = 0.1, T_B = 0.2$

図 4-31 波形誤差法による van der Pol 方程式の学習曲線



(a) T_B と T_C との関係

◇ : $\eta = 0.10$, + : $\eta = 0.20$, □ : $\eta = 0.40$, × : $\eta = 0.80$



(b) η と T_C との関係

◇ : $T_B = 0.1$, + : $T_B = 0.2$, □ : $T_B = 0.4$, × : $T_B = 0.8$

図 4-32 波形誤差法における学習パラメータと学習時間との関係

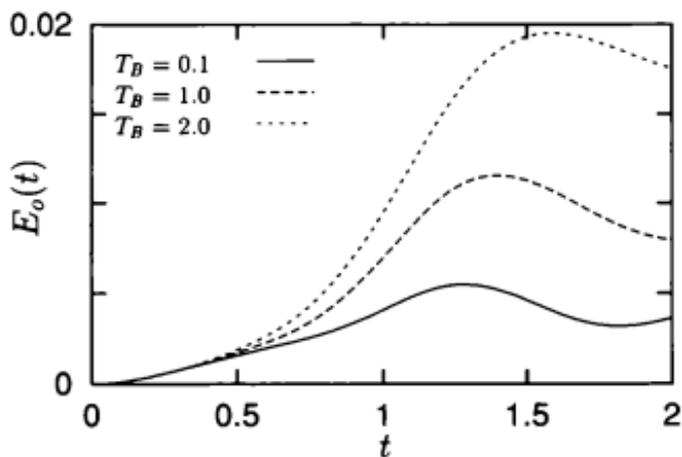


図 4-33 速度誤差法の軌道追従能力の比較

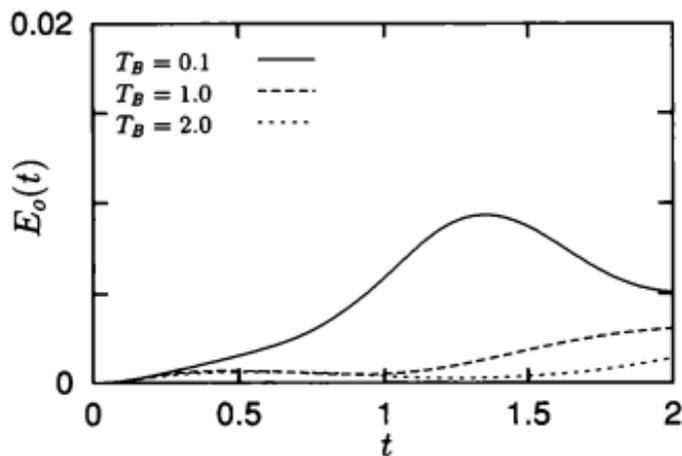


図 4-34 波形誤差法の軌道追従能力の比較

図4-33は、 $T_B = 0.1, 0.2, 2.0$ の三とおりの条件で速度誤差法によって学習したRNNの軌道追従能力を比べており、 T_B が大きいほど追従性が悪く、誤差の増加が速いことがわかる。速度誤差法の場合はRNNの出力を教師信号に拘束したまま学習をしているので、 T_B を大きくしても軌道の追従能力は向上しない。むしろ第3章の式(3.27)での実質的な学習係数 $T_B \cdot \eta$ が大きくなるため、学習結果が安定しなくなって追従性が悪化するものと考えられる。

一方、波形誤差法の場合は、図4-34に示すように、 T_B が大きいほど誤差の増加は速く軌道追従性がよくなる。また、同一の T_B で学習した場合には、波形誤差法のほうが速度誤差法より追従性がよい。これは、波形誤差法では、軌道の誤差を評価関数として、それを小さくするように学習しているのだから当然とも言えるが、波形誤差法の場合は関数誤差を小さくし過ぎるという過学習を自然に回避する能力があるように思われる。

4.4.2 Rössler 方程式の学習

Rössler 方程式

$$\begin{aligned} \frac{dq_1}{dt} &= -q_2 - q_3, \\ \frac{dq_2}{dt} &= q_1 + aq_2, \\ \frac{dq_3}{dt} &= b + q_1q_3 - cq_3 \end{aligned} \quad (4.14)$$

は、パラメータが $a = 0.2, b = 0.2, c = 5.7$ のとき、図4-35に示したカオスアトラクタ(Rösslerアトラクタ)を持つ。このRösslerアトラクタを教師信号として学習実験を行なった。微分方程式の求解にはvan der Pol方程式の学習の場合と同じくルンゲクッタ法を用い、刻み幅は $h = 0.05$ とした。

図4-36は速度誤差法による学習回数10万回の学習曲線で、学習の条件はシグモイドユニット数 $M = 50$ 、学習係数 $\eta = 0.01$ 、 $T_B = 0.5$ である。学習に伴い、誤差は減少しているが、学習後のRNNが出力したアトラクタは図4-37の周期アトラクタであり、カオスの学習には成功していない。

図4-38は波形誤差法による学習回数10万回の学習曲線で、学習の条件は速度誤差法の場合と同じである。図4-36の速度誤差法の場合に比べて学習は不安定で、最終的な誤差も大きい。学習後のRNNは図4-39に示すような、Rösslerアトラクタと定性的には同じカオスアトラクタを出力した。

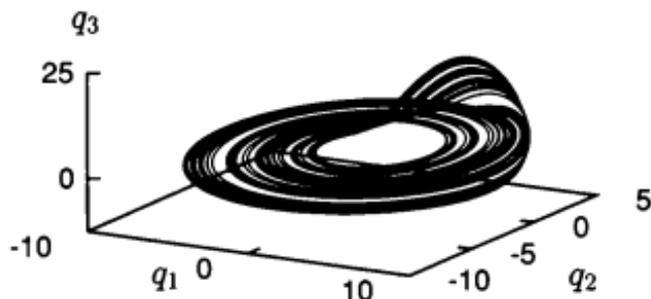


図4-35 Rössler アトラクタ ($a = 0.2, b = 0.2, c = 5.7$)

二乗関数誤差の大きい波形誤差法によって学習したRNNの方が、「アトラクタの学習」という意味で速度誤差法より優れているのは、前述のように、波形誤差法が関数を近似しすぎることなく、カオスを生成するという系の特徴を学習する能力を持っているためと考えられる。また、学習後のRNNによる教師信号の軌道の追従能力をvan der Pol方程式の学習の場合と同様に計算し、図4-40に示した。軌道の追従能力の点でも、波形誤差法が速度誤差法より優れていることがわかる。

4.5 結言

本章では、1次元離散力学系および2次元、3次元連続力学系を対象とした、数値シミュレーションによる学習実験結果について考察した。本章で得られた成果は以下のように要約される。

- (1) 離散力学系の学習において、学習係数 η 、学習回数 N_B の関数誤差法は学習初期の学習速度で評価するかがり、学習係数 $N_B \cdot \eta$ のBP法と等価であることを数値的に示し、第3章で展開した確率的降下法に基づく議論を検証した。
- (2) 軌道誤差法による1次元離散系カオスの学習過程において学習が急激に加速さ

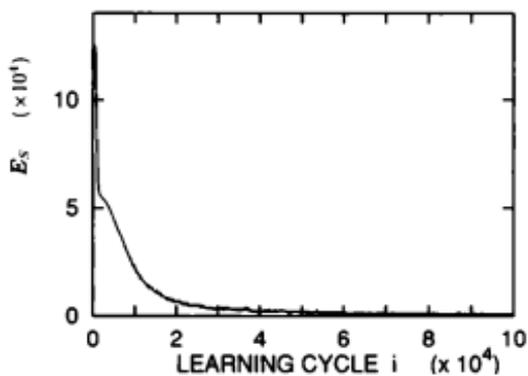


図 4-36 速度誤差法による学習曲線

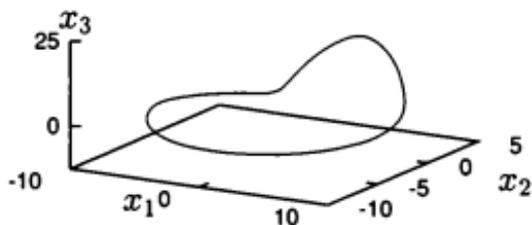


図 4-37 速度誤差法による学習結果のアトラクタ

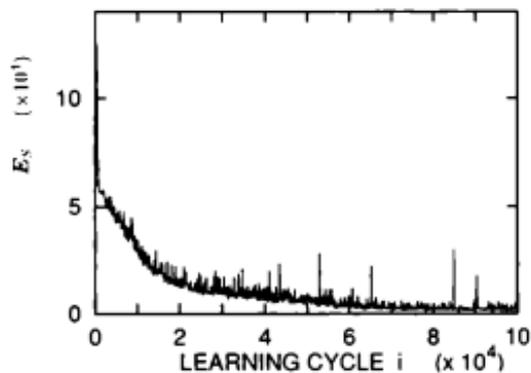


図 4-38 波形誤差法による学習曲線

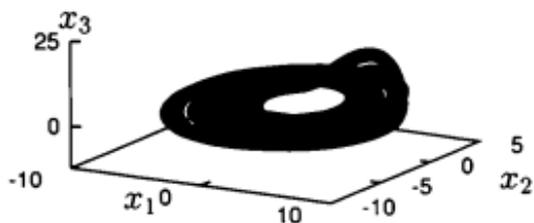
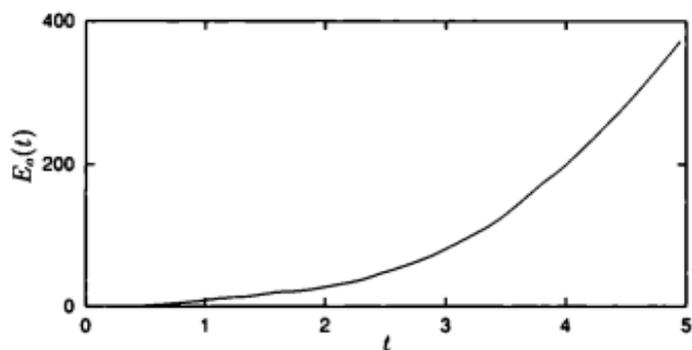
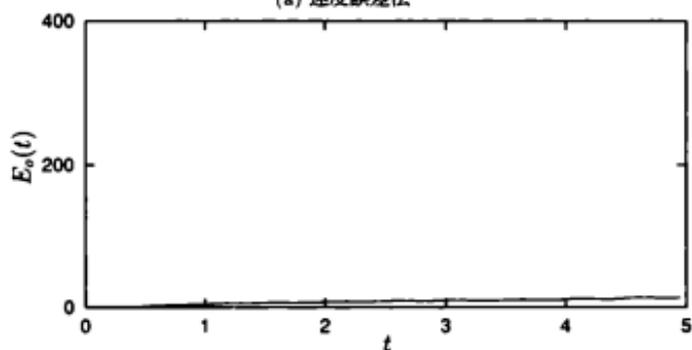


図 4-39 波形誤差法による学習結果のアトラクタ



(a) 速度誤差法



(b) 波形誤差法

図 4-40 軌道追従能力の比較

れる現象を発見し、これが系の対称性と逐次型の勾配降下法が本質的に持つ揺らぎおよび学習されたカオスの揺らぎに起因することを明らかにした。

(3) 離散力学系の学習における教師信号の提示方法について検討し、以下の結果を得た。まず、有限個の時系列データを繰り返し提示して学習させた場合、データ数が20程度でデータ数が無制限の場合と同程度の速度で学習が行なえることを示した。次に、Cubic map の学習の場合の軌道誤差法による加速現象はデータ数が数十の段階で生じていることを明らかにした。また、教師信号の分布密度は学習速度に大きく影響し、これに比べると時間相関が及ぼす影響は小さいこと、一般に一様白色雑音で教師パターンを生成すると学習は遅くなることを示した。

(4) 連続力学系の学習においても、学習係数が小さい場合、学習初期の速度は確率的降下法の理論に基づく評価式に従うことを確認した。

(5) 連続力学系のカオスの学習では、速度誤差法より波形誤差法による学習結果のほうが軌道追従性、アトラクタの近似の点で優れていることを示した。

確率的降下法の理論により逐次型勾配降下法による学習の平均的な進行状況を把握できる。特に、周期軌道の学習や学習の初期段階においては、一括型学習とほぼ等価であることが数値的に示された。しかし、カオスを学習する場合は、カオスの有する揺らぎの性質が学習の最終段階で大きく影響し、確率的降下法の理論からは予測できない現象を呈することがある。

本章で得られた結果は、時系列の学習とはその時系列を生成する系を表わす関数の近似にすぎないとしても、その学習過程において時系列の性質は重要な役割を演じること、特に時系列がカオスである場合にその動きは顕著であることを示している。

第5章 差分方程式の理論を用いた学習収束過程の解析

5.1 緒言

本章では、力学系の学習の最適状態への収束条件を非自律的差分方程式の理論を用いて解析する。

前章までに考察した、離散力学系の学習および最適制御型学習則を用いた連続力学系の学習は、教師信号とRNNの出力の誤差を E とし、結合パラメータを w と表わすと、一般に次のような形式の勾配降下法として定式化される。

$$w(k+1) = w(k) - \eta \nabla_w E(w(k), k). \quad (5.1)$$

ここで、 η は学習係数、 k は学習サイクルを表わす離散時間である。

式(5.1)は、パラメータ w に関する差分方程式(離散力学系)である。さらに、逐次型の学習則では一般に誤差評価関数 E は時間 k に陽に依存するので、この方程式は非自律的である。

本章では、力学系の学習則(5.1)を結合パラメータ w を変数とし、教師信号によって駆動される非自律的差分方程式と考える。このように学習を力学系とみなした場合、学習の目標である最適状態は力学系の不動点となる。そこで、最適点近傍での学習の収束条件を、力学系の不動点の安定条件として考察する。

第3章では、逐次修正型の学習アルゴリズムによる離散力学系の学習過程を確率的降下法の理論を用いて解析し、これらのアルゴリズムは平均的な意味で逐次型のBP法に等価であることを示した。そこでは力学系の不変測度の存在とその不変測度に関するエルゴード性を仮定し、学習する時系列をランダム系列として扱った。しかし、力学系から生成される時系列はカオスも含めランダムではなく、あくまで決定論的法則にしたがっている。したがって、学習は確定的な時系列によって駆動された力学系と考えることができ、その解析には差分方程式の不動点の安定性理論が適用できることになる。

このような差分方程式の不動点の安定問題としての学習過程の解析は甘利によりなされており、平均的な収束速度や最適状態近傍でのパラメータの共分散行列が理論的に導出されている^{43,44)}。本章では、この成果を補足しつつ、カオス力学系

理論に基いた新たな観点から解析を行ない、学習対象の時系列が、周期的、概周期的、カオス的な場合のそれぞれに対して最適状態の局所収束条件を導出する。

5.2 離散力学系の学習の収束性

5.2.1 学習を表わす差分方程式

次式で表わされる m 次元離散力学系を考える。

$$q(k+1) = f(q(k)) \quad (q \in R^m). \quad (5.2)$$

f は R^m の上の連続写像であり、 k は離散時間を表わす。各時刻 k において、 $q(k)$ を入力、 $q(k+1) = f(q(k))$ を出力とするパターンを教師信号として与え、この系を次式で記述される RNN で学習する。

$$x(k+1) = \tilde{f}(x(k), w) \quad (x \in R^m). \quad (5.3)$$

w は結合パラメータを表わす n 次元ベクトル ($w \in R^n$) である。

BP法による逐次型学習則は

$$w(k+1) = w(k) - \eta \nabla_w E(w(k), k) \quad (5.4)$$

で記述され、誤差評価関数は

$$E(w, k) = \frac{1}{2} \|f(q(k)) - \tilde{f}(q(k), w)\|^2 \quad (5.5)$$

である。教師信号 $q(k)$ の値は各時刻 k において異なるから、誤差評価関数 E は k に陽に依存する。したがって、学習を記述する式 (5.4) はパラメータ w を変数とする非自律的差分方程式になり、すべての時刻 k で $E(w, k) = 0$ となる最適点 $w = w^*$ は、この差分方程式の不動点である。 w^* が方程式 (5.5) の不動点として漸近安定であれば、学習が最適状態に局所的に収束することになる。

同様に、関数誤差法（一括型 BP 法）および軌道誤差法による学習則は、

$$w(k+N_B) = w(k) - \eta \nabla_w E(w(k), k) \quad (5.6)$$

で記述され、誤差評価関数はそれぞれ次のように定義される。

$$E(w, k) = \frac{1}{2} \sum_{l=0}^{N_B-1} \|f(q(k+l)) - \tilde{f}(q(k+l), w)\|^2, \quad (5.7)$$

$$E(w, k) = \frac{1}{2} \sum_{l=0}^{N_B-1} \|f^{l+1}(q(k)) - \tilde{f}^{l+1}(q(k), w)\|^2. \quad (5.8)$$

5.2.2 二乗誤差を最小化する勾配降下法の収束条件

逐次型BP法による離散力学系の学習

まず、BP法による学習を、誤差評価関数が複数の関数の二乗和として表わされるような勾配降下法として一般化し、その収束性を考察する。

誤差評価関数 E がパラメータ w と離散時間 k の関数 m 個の二乗和として表現される場合の勾配降下法を考える。すなわち、式(5.5)を一般化し、

$$E(w, k) = \frac{1}{2} \sum_{i=1}^m F_i(w, k)^2 \quad (5.9)$$

として

$$w(k+1) = w(k) - \eta \nabla_w E(w(k), k) \quad (5.10)$$

の更新則により、すべての整数 k について

$$F_i(w^*, k) = 0 \quad (1 \leq i \leq m) \quad (5.11)$$

となるような最適点 w^* を探索するような勾配降下法である。

以下、このような最適点 w^* が少なくとも一つ存在するものとして、力学系(5.10)の不動点としてのその安定性を考察する。

まず、式(5.10)を

$$w(k+1) = g(w(k), k) = (g_1(w(k), k), \dots, g_n(w(k), k))^T$$

と表わすことにし (τ は行列の転置を表わす)、 w^* の近傍での線形近似系、

$$w(k+1) = Dg(w^*, k)w(k) \quad (5.12)$$

の零解 $w(k) = 0$ の安定性を考える。

$$g_i(w(k), k) = w_i - \eta \frac{\partial E(w(k), k)}{\partial w_i} \quad (1 \leq i \leq n) \quad (5.13)$$

であるから、式(5.12)において関数 g のヤコビ行列 Dg は、

$$\frac{\partial g_i(w(k), k)}{\partial w_j} = \delta_{ij} - \eta \frac{\partial^2 E(w(k), k)}{\partial w_i \partial w_j} \quad (1 \leq i, j \leq n) \quad (5.14)$$

となる。 δ_{ij} はクロネッカーのデルタである。特に最適点 $w(k) = w^*$ においては、式(5.9)と(5.11)より、

$$\frac{\partial g_i(w^*, k)}{\partial w_j} = \delta_{ij} - \eta \sum_{l=1}^m \frac{\partial F_l(w^*, k)}{\partial w_i} \frac{\partial F_l(w^*, k)}{\partial w_j} \quad (1 \leq i, j \leq n) \quad (5.15)$$

である。ここで、

$$\mathbf{a}^{(i)}(k) = (a_{k1}^{(i)}, \dots, a_{kn}^{(i)})^r, \quad (5.16)$$

$$a_{ki}^{(i)} = \frac{\partial F_i(\mathbf{w}^*, k)}{\partial w_i} \quad (1 \leq i \leq m, 1 \leq i \leq n, k \in \mathcal{Z})$$

とおけば、 \mathbf{w}^* での関数 g のヤコビ行列は、

$$Dg(\mathbf{w}^*, k) = I - \eta \sum_{i=1}^m \mathbf{a}^{(i)}(k) \cdot \mathbf{a}^{(i)}(k)^r \quad (5.17)$$

と表わされる。ここで、 I は n 次単位行列である。式 (5.17) は、また、

$$Dg(\mathbf{w}^*, k) = \frac{1}{m} \sum_{i=1}^m A^{(i)}(k), \quad (5.18)$$

$$A^{(i)}(k) = I - m\eta \mathbf{a}^{(i)}(k) \cdot \mathbf{a}^{(i)}(k)^r \quad (5.19)$$

と表わせる。

式 (5.19) の行列 $A^{(i)}(k)$ は次の性質を持つことがわかる。

[性質] 任意の整数 k および任意の i ($1 \leq i \leq m$) について $A^{(i)}(k)$ は固有値

$$\lambda^{(i)}(k) = 1 - m\eta \|\mathbf{a}^{(i)}(k)\|^2 \quad (5.20)$$

の固有ベクトル $\mathbf{a}^{(i)}(k)$ と、これに直交する固有値 1 の $(n-1)$ 次元の固有空間 $V_k^{(i)}$ を持つ。

(証明) 式 (5.19) より、

$$\begin{aligned} A^{(i)}(k) \cdot \mathbf{a}^{(i)}(k) &= \mathbf{a}^{(i)}(k) - m\eta \mathbf{a}^{(i)}(k) \|\mathbf{a}^{(i)}(k)\|^2 \\ &= (1 - m\eta \|\mathbf{a}^{(i)}(k)\|^2) \mathbf{a}^{(i)}(k) \end{aligned}$$

であるから、前半の成立することがわかる。また、 $\mathbf{a}^{(i)}(k)$ に直交する任意のベクトル \mathbf{v} について、

$$A^{(i)}(k) \cdot \mathbf{v} = \mathbf{v} - m\eta \mathbf{a}^{(i)}(k) (\mathbf{a}^{(i)}(k) \cdot \mathbf{v}) = \mathbf{v}$$

であるから、後半が成立する。 □

以下、関数 F_i が時間 k に関して、(1) 周期的、(2) 概周期的、(3) カオス的な場合に分けて考察を進める。これらは、それぞれ、周期軌道、概周期軌道、カオス軌道の B P

学習の一般化である。したがって、 F_i の周期性は i に依存しないものと仮定しておく。

(1) 関数 F_i が時間 k に関して周期的な場合

まず、次の定理が成り立つ。

【定理 5-1】 関数 $F_i(w, k)$ ($1 \leq i \leq m$) が k に関して周期 M を持ち、

$$\|a^{(i)}(k)\| \neq 0 \quad (1 \leq i \leq m, 1 \leq \forall k \leq M)$$

であるとし、

$$|\lambda^{(i)}(k)| = |1 - m\eta \|a^{(i)}(k)\|^2| < 1 \quad (1 \leq i \leq m, 1 \leq \forall k \leq M)$$

となるよう η を十分小さく選んでおくとする。このとき、

$$B = Dg(w^*, M) \cdots Dg(w^*, 1)$$

とすると、

(i) $mM < n$ の場合

B は固有値 1 を持ち、そのノルムは $\|B\| = 1$ である。

(ii) $mM \geq n$ の場合

$n \times mM$ 行列 A を、

$$A = (a^{(1)}(1), \dots, a^{(m)}(1), \dots, a^{(1)}(M), \dots, a^{(m)}(M))$$

と定義すると、

$$\|B\| < 1 \Leftrightarrow \text{rank} A = n.$$

((i) の証明) まず、各 $A^{(i)}(k)$ は上述の性質を持つこと、すなわち絶対値が 1 に満たない固有値 $\lambda^{(i)}(k)$ の固有ベクトルと、それに直交する固有値 1 の $(n-1)$ 次元の固有空間 $V_k^{(i)}$ を持つことから、 $\|A^{(i)}(k)\| = 1$ である。したがって、式(5.18)のようにこれらの行列の平均として表わされるヤコビ行列 $Dg(w^*, k)$ のノルムは $\|Dg(w^*, k)\| \leq 1$ を満たす。したがって、その積である行列 B についても $\|B\| \leq 1$ であることがわかる。

また、 $\dim(V_k^{(i)}) = n-1$ の関係と $mM < n$ の仮定より、 $\dim(\cap_{k=1}^M \cap_{i=1}^m V_k^{(i)}) \geq 1$ である。したがって B は $\cap_{k=1}^M \cap_{i=1}^m V_k^{(i)}$ の中に固有値 1 の固有ベクトルを持つ。よって $\|B\| = 1$ である。 □

(ii)の証明)

(\Rightarrow) $mM \geq n$ であるから $\text{rank} A \leq n$ である。もし $\text{rank} A < n$ ならば、 $\dim(\cap_{k=1}^M \cap_{i=1}^m V_k^{(i)}) \geq 1$ である。なぜならば、 $\dim(\cap_{k=1}^M \cap_{i=1}^m V_k^{(i)}) = 0$ であれば、 $\alpha^{(1)}(1), \dots, \alpha^{(m)}(M)$ は $V_1^{(1)}, \dots, V_M^{(m)}$ とそれぞれ直交するから、これらの mM 個のベクトルの中に線形独立な n 個の組が存在するので、 $\text{rank} A = n$ となるからである。したがって、 $\text{rank} A < n$ を仮定すると、 B は $\cap_{k=1}^M \cap_{i=1}^m V_k^{(i)}$ の中に固有値 1 の固有ベクトルを持ち、 $\|B\| \geq 1$ となって $\|B\| < 1$ に矛盾する。よって $\text{rank} A = n$ である。

(\Leftarrow) (i) と同様、 $\|B\| \leq 1$ が成り立つことが容易にわかる。もし $\|B\| = 1$ であれば、 $\|Bv\| = \|v\|$ となる非零ベクトル v が存在することになるが、各 $A^{(i)}(k)$ はノルムが 1 以下であるから $v \in \cap_{k=1}^M \cap_{i=1}^m V_k^{(i)}$ である。ところが、 $\text{rank} A = n$ であるから $\dim(\cap_{k=1}^M \cap_{i=1}^m V_k^{(i)}) = 0$ なので、矛盾する。よって、 $\|B\| < 1$ である。 \square

定理 5-1 の (i) は、BP による学習の観点からは次のように解釈できる。すなわち、教師信号の次元と周期の積がパラメータの個数より少ない場合は、最適値の集合を

$$W_{OP} = \{ w \mid E(w, k) = 0, 1 \leq k \leq M \}$$

と表わすと、この集合は w^* で孤立しておらず、一般には w^* を含む $\dim(\cap_{k=1}^M \cap_{i=1}^m V_k^{(i)})$ 次元の多様体になっている。このとき (5.10) の離散系の不動点 w^* は漸近安定ではないが、反復写像は W_{OP} 上に収束することが期待される。

また定理 5-1 の (ii) は、行列 A の階数が n であれば、すなわち W_{OP} の点 w^* が孤立していれば、(5.10) の反復写像は w^* の近傍で縮小的になっていることを意味する。このことから、後に示すように w^* が漸近安定であることが導かれる。

式 (5.10) の不動点の漸近安定性を示すために、線形化原理あるいはリアプノフの第一の方法⁴⁵⁾ に基づく次の補題を述べておく。

[補題 5-1] ¹¹⁾ 離散力学系

$$w(k+1) = g(w(k), k) \quad (w \in R^n) \quad (5.21)$$

の不動点を w^* とすると、その近傍での線形近似系

$$w(k+1) = Dg(w^*, k)w(k) \quad (5.22)$$

の零解が一様漸近安定ならば、 w^* も一様漸近安定になる。 \square

ここで、(5.21)の解 w^* が一様漸近安定とは次のことを意味する。すなわち、初期時刻 $k = k_0$ で値 w_0 をとる解を $w(w_0, k_0, k)$ と表わすと、次の二つの条件を満たすとき w^* は一様漸近安定であるという¹³⁾。

(i) 任意の正数 ϵ に対し正数 δ が存在して、 $\|w_0 - w^*\| < \delta$ ならば、任意の整数 k_0 とすべての $k \geq k_0$ について $\|w(w_0, k_0, k) - w^*\| < \epsilon$ となる (一様安定性)。

(ii) 正数 δ があって、任意の正数 ϵ に対し整数 $K(\epsilon)$ が存在して、 $\|w_0 - w^*\| < \delta$ ならば、任意の整数 k_0 とすべての $k \geq k_0 + K(\epsilon)$ について $\|w(w_0, k_0, k) - w^*\| < \epsilon$ となる。

そこで、この補題を利用して式(5.10)の不動点の漸近安定性を示す。

[定理5-2] 離散力学系(5.10)の不動点 w^* は定理5-1の条件および次の二つの条件を満たすとき一様漸近安定になる。

(条件1) 関数 $F_i(w, k)$ ($i = 1, \dots, m$) は k に関して周期 M を持ち、 $mM \geq n$ である。

(条件2) $\text{rank} A = n$ である。ここで、 A は定理5-1で定義した $n \times mM$ 行列である。

(証明) 補題5-1より、線形系(5.12)の零解の一様漸近安定性を示せば十分である。

定理5-1の証明で述べたように、任意の k について $\|Dg(w^*, k)\| \leq 1$ である。したがって、一様漸近安定性の条件(i)が線形系(5.12)の零解に関して成り立つことは明らかである。

また、定理5-1と条件2より、

$$\alpha = \|B\| < 1$$

となる実数 α が存在することがわかる。そこで、 $\delta = 1$ として、任意の正数 ϵ に対し、 $\alpha^{K(\epsilon)/M} < \epsilon$ となるような $K(\epsilon)$ を選ぶと、 $w_0 < \delta$ ならば、任意の整数 k_0 とすべての $k \geq k_0 + K(\epsilon)$ について $\|w(w_0, k_0, k)\| < \epsilon$ となる。よって、線形系(5.12)の零解は一様漸近安定になり、補題5-1より、系(5.10)の不動点 w^* も一様漸近安定になる。□

定理5-2の条件は、 mM 個 ($M \geq n$) のベクトル $a^{(1)}(1), \dots, a^{(m)}(M)$ のうちの n 個が線形独立であることを意味するので、一般的 (generic) な仮定である。したがって、周期軌道の周期と学習対象の次元の積が RNN の結合パラメータの総数より大きい場合の逐次型BP法による学習は、最適値の十分近くに初期状態があれば一般的に収束することが示せたことになる。

(2) 関数 F_i が時間 k に関して概周期的な場合

この場合は、逐次型BP法による概周期軌道の学習を想定して、関数 F_i の時間

項が m 次元離散力学系、

$$q(k+1) = f(q(k)) \quad (q \in \mathbb{R}^m)$$

から生成される概周期軌道 $q(k)$ によって定まるものとし、式 (5.10) を

$$w(k+1) = \tilde{g}(w(k), q(k)) \tag{5.23}$$

と表わす。

ここで、離散時系列 $q(k)$ が概周期的であるとは、次のことを意味する。すなわち、任意の $\epsilon > 0$ に対し正整数 L が存在し、任意の整数 k, l について $0 \leq j \leq L-1$ となる j が存在して、

$$\|q(k) - q(k+l+j)\| < \epsilon$$

となる^{12, 13)}。 $q(k)$ が周期 M を持つ場合は $L = M$ とおけば、任意の ϵ について上の条件が成立する。したがって、周期軌道はすべて概周期軌道である。しかし、周期軌道の学習については既に考察したので、ここでは $q(k)$ は非周期的であることを仮定する。

このとき、次の定理が成立する。

[定理 5-3] 軌道 $q(k)$ が概周期的 (かつ非周期的) であれば、(5.23) の不動点 w^* は一般に一様漸近安定になる。

(証明) 定理 5-2 と同様、(5.23) の線形近似系

$$w(k+1) = D\tilde{g}(w^*, q(k))w(k) \tag{5.24}$$

の一様漸近安定性を示し、補題 5-1 によって (5.23) の不動点の一様漸近安定性を証明する。まず、

$$B(l) = D\tilde{g}(w^*, q(l+n-1)) \cdots D\tilde{g}(w^*, q(l))$$

と定義する。仮定より $q(k)$ は非周期的であるから、定理 5-1 の証明と同様に、

$$\|B(l)\| < \beta(l) < 1$$

となる $\beta(l)$ が存在する (一般に $\dim(\cap_{i=1}^n \cap_{j=1}^m V_i^{(i)}) = 0$ である)。また、 $q(k)$ の概周期性と $A^{(i)}(k)$ の $q(k)$ に関する連続性より、正整数 L が存在して、任意の整数 l に対し $0 \leq j \leq L-1$ となる j が存在して、

$$\| \|B(0)\| - \|B(l+j)\| \| < \frac{1 - \beta(0)}{3}$$

となる。よって

$$\|B(l+j)\| < \frac{1+2J(0)}{3} < 1$$

である。ここで、 L はいくらでも大きく選べるから $L > n$ とし、また、 $J = (1+2J(0))/3$ とすると、任意の整数 l について、

$$\|Dg(w^*, q(l+2L-1)) \cdots Dg(w^*, q(l))\| < J < 1$$

である。なぜならば、 $l \leq l+j < l+j+n-1 < l+2L-1$ であり、また任意の k について $\|Dg(w^*, q(k))\| \leq 1$ であるから。

したがって、定理 5-2 の証明と同様に線形近似系 (5.24) の一様漸近安定性が成立し、(5.23) の不動点 w^* も一様漸近安定となる。□

(3) 関数 F_t が時間 t に関してカオス的な場合

概周期的な場合と同様、この場合は、逐次型 B P 法によるカオス軌道の学習を想定して系 (5.23) の不動点 w^* の安定性を考える。

カオスはその定義自体が明確になっていないが、ここでは、少なくとも次の (a)、(b)、(c) の三つの性質を持つものと仮定する⁴⁾。

$q(k)$ をカオス軌道とし、その閉包を

$$A_T = \overline{\{q(k) \mid -\infty < k < +\infty\}}$$

とすると、

(a) A_T はコンパクト、(b) $q(k)$ は非周期的、(c) A_T は不安定な周期軌道を含む。

性質 (a) は定常状態 (アトラクタ) のみを考えることを意味する。性質 (b) は概周期軌道でも有するが、(c) はカオスに特有の性質であり、以下の議論で重要な役割を演じる。

カオスの場合は周期的な場合や概周期的な場合と異なり、線形近似系の零解の一様漸近安定性から元の非線形系 (5.23) の不動点の漸近安定性を示すことは困難である。実際、ある条件のもとで線形近似系の零解が一様漸近安定にはならないことがわかる。このことを以下で示す。

⁴⁾ これは、カオスアトラクタの非極小性ということもできる。すなわち、概周期軌道の閉包はコンパクトならば極小集合であるが¹²⁾、カオス軌道の閉包は不安定な周期軌道を含むため、極小的ではない³⁰⁾。ただし、極小集合とは、空でない閉不変集合であり、空でない閉不変集合を真部分集合として持たないものをいう。

[定理5-4] $q(k)$ を周期 n/m 未満の周期点を少なくとも一つ含むカオスアトラクタ A_T 上で稠密な軌道とする。このとき、線形近似系(5.24)の零解は一様漸近安定ではない。

(証明) $\hat{q}(k)$ を A_T に含まれる周期 $l (< n/m)$ の周期軌道とし、その軌道上の1点を \hat{q}_0 とすると、定理5-1の(i)より、任意の k に対し、

$$\|D\hat{g}(w^*, \hat{q}(k+l)) \cdots D\hat{g}(w^*, \hat{q}(k+1))\| = 1$$

である (q は A_T 上で稠密だから、(5.23) 式の q を \hat{q} に換えても $F_t(w^*, k) = 0$ が成り立つ)。したがって、 $D\hat{g}$ とノルムの連続性および $q(k)$ を生成する力学系の初期値に対する連続性より、任意の整数 K に対し正数 β が存在して、任意の整数 k_0 に対し $\|q(k_0) - \hat{q}_0\| < \beta$ ならば、

$$\|D\hat{g}(w^*, q(k_0+K)) \cdots D\hat{g}(w^*, q(k_0+1))\| > \frac{1}{2}$$

となる。仮定より $q(k)$ は A_T で稠密だから $\|q(k_0) - \hat{q}_0\| < \beta$ となる k_0 は必ず存在する。したがって、任意の正数 δ に対し $\epsilon = (1/4)\delta$ とすると、任意の K に対し k_0 と $\|w_0\| < \delta$ となる w_0 が存在して ($\|w_0\| > (1/2)\delta$ となるように選ぶことができる)、

$$\|D\hat{g}(w^*, q(k_0+K)) \cdots D\hat{g}(w^*, q(k_0+1))w_0\| > \epsilon$$

となる。したがって(5.24)の零解は一様漸近安定ではない。 □

この定理より、カオス学習の収束性を線形近似系の安定性から議論することは困難であることがわかる。

ただし、ある条件のもとでは線形近似系(5.24)の零解が漸近安定であることは以下のように示せる。

[定理5-5] 時系列 $\{q(k)\}$ がカオスアトラクタ A_T に含まれ、正整数 n' が存在して、 A_T の不変確率測度 P に関して $f^{n'}$ はエルゴード的で、

$$P(\{q \mid r(q) < 1\}) > 0 \tag{5.25}$$

であるとする。ただし、

$$r(q) = \|D\hat{g}(w^*, f^{n'-1}(q)) \cdots D\hat{g}(w^*, q)\|$$

とする。さらに、すべての k と j について $|\lambda^{(j)}(k)| < 1$ となるように式(5.20)の η を十分小さく選んでおく。このとき、(5.24)の零解が漸近安定であることが $q(k)$ の初期値に関して確率1で成り立つ。

(証明) まず、仮定より $P(\{q \mid r(q) < a\}) > \gamma$ となる正数 $a (< 1)$ と $\gamma (< 1)$ が存在することがわかる。

また、 P に関して f^n はエルゴード的であるから、集合 $\{q \mid r(q) < a\}$ の特性関数を $\xi(q)$ とすると、

$$\begin{aligned} \lim_{l \rightarrow \infty} \frac{1}{l} \sum_{i=0}^{l-1} \xi(f^{n^i}(q)) &= \int \xi(q) dP \\ &= P(\{q \mid r(q) < a\}) > \gamma \quad (P\text{-a.e.}). \end{aligned}$$

したがって、 P に関してほとんどすべての q に対し整数 $m(q)$ が存在して、 $l \geq m(q)$ ならば、

$$\sum_{i=0}^{l-1} \xi(f^{n^i}(q)) > l\gamma$$

である。よって、

$$r(f^{n^i}(q)) < a$$

となる $i = 0, \dots, l-1$ の個数は $l\gamma$ 以上である。また、 $|\lambda^{(j)}(k)| < 1$ の仮定より $\|Dg(w^*, f^k(q))\| \leq 1$ であるから、

$$\|Dg(w^*, f^{n^{l-1}}(q)) \cdots Dg(w^*, q)\| \leq \prod_{i=0}^{l-1} r(f^{n^i}(q)) < a^{l\gamma}$$

となる。

そこで、 $\delta = 1$ とし、任意の正数 ϵ に対して $a^{\gamma m(q)} < \epsilon$ となるように $m(q)$ を十分大きく選び、 $K(\epsilon, q) = m(q)n^l$ とすると、 $k \geq K(\epsilon, q)$ ならば、 $\|w_0\| < \delta$ となる任意の w_0 に対して、

$$\|Dg(w^*, f^k(q)) \cdots Dg(w^*, q)w_0\| \leq a^{\gamma m(q)} < \epsilon$$

となる。よって(5.24)の零解は漸近安定である。 □

また、この定理の系として次が導かれる。

[系] 式(5.23)において、関数 g がパラメータ w に関してアフィンであり、かつ定理5-1の条件が成り立てば、式(5.23)の不動点 w^* は時系列 $q(k)$ の初期値に関して確

率1で漸近安定になる。

□

定理5-5の条件のうち、式(5.25)は定理5-1の(ii)の行列 A が $M = n'$ として $\text{rank} A = n$ であるような $q(k)$ 、すなわち、 $n'm$ 個のベクトル $a^{(1)}(1), \dots, a^{(m)}(n')$ が線形独立であるような $q(k)$ の初期値が正の測度を持つことを意味する。したがって、 $n' \geq n/m$ であればこれは一般的に成立することがわかる。また、 f^n のエルゴード性は f が混合的であれば成立する。以上より、式(5.24)の不動点は、ごく一般的に漸近安定であることがわかる。

また、定理5-5の証明より、 P に関してほとんどすべての q に対し $K > 0$ が存在して $k \geq K$ ならば、

$$\begin{aligned} \frac{1}{k} \log \left\| \prod_{i=0}^{k-1} Dg(w^*, f^i(q)) \right\| &\leq \frac{\gamma m(q)}{K} \log \alpha \\ &\leq \frac{\gamma}{n'} \log \alpha < 0 \end{aligned}$$

である(ここで、 Π は行列 $Dg(w^*, f^i(q))$ を左からかけていくことを表す)。したがって、定理5-5の条件が満たされれば、 g の最大法線リアプノフ指数⁴⁶⁾

$$\mu = \lim_{k \rightarrow \infty} \frac{1}{k} \log \left\| \prod_{i=0}^{k-1} Dg(w^*, f^i(q)) w \right\| \quad (5.26)$$

は負であることがわかる。法線リアプノフ指数は、式(5.23)と m 次元離散系 $q(k+1) = f(q(k))$ とを合わせた $(n+m)$ 次元自律系方程式のカオス的不变集合 $(w^*, q(k))$ の安定性を評価する量であり、カオスの同期条件の評価に利用されている^{46, 47, 48)}。最大法線リアプノフ指数が負であれば、線形系(5.24)の零解の漸近安定性は保証されるが、定理5-4で示したように一様漸近安定ではない。つまり、式(5.26)の収束の速度が初期値 q に依存するのである。

1次元時系列データからの m 次元離散力学系の学習

以上の力学系学習の定式化は、 m 次元力学系のすべての状態変数が観測可能で、教師信号として m 次元ベクトルの組 $(q(k), q(k+1))$ が提示されることを前提としていた。しかし、現実の時系列学習では、系の状態変数がすべて観測されることは稀であり、一つの状態変数すなわちスカラー値の時系列のみが観測可能であることが多い。このような条件での力学系の学習法としては、適当な正整数 r と L を選定し、スカラー値時系列 $q_1(k)$ から L 次元ベクトル値時系列、

$$(q_1(k), q_1(k-r), \dots, q_1(k-(L-1)r))$$

を構成して、これを教師信号として RNN に学習させる方法がとられている^{19, 21)}。ここで、 $q_1(k)$ が式 (5.2) の m 次元力学系の m 変数の一つであるとする²²⁾、 $L \geq 2m+1$ ならば、構成した L 次元時系列は式 (5.2) と力学論的に等価な系に一般的に従うことが保証されている⁶⁰⁾。このような場合の RNN は、

$$x(k+1) = \hat{f}(x(k), x(k-\tau), \dots, x(k-(L-1)\tau), w) \quad (5.27)$$

で記述され ($x \in R$)、誤差評価関数は

$$E(w, k) = \frac{1}{2} F_1(w, k)^2 \quad (5.28)$$

$$F_1(w, k) = f(q_1(k)) - \hat{f}(q_1(k), q_1(k-\tau), \dots, q_1(k-(L-1)\tau), w)$$

と定義される。この誤差評価関数は BP 法による 1 次元系の学習と同じく関数 F_1 の二乗で表わされる。したがって 1 次元系の学習の場合と全く同様に、定理 5-1~定理 5-5 の結果が成立する。すなわち、RNN のパラメータ数より高い次数の周期軌道、一般の概周期軌道、およびパラメータが線形の RNN でカオスを学習する場合には、一般的な仮定のもとに学習が局所的に収束する。

関数誤差法による学習

関数誤差法による学習則は式 (5.6) と (5.7) で定義される。また、

$$F_i(w(k), k, l) = \hat{f}_i(q(k+l), w(k)) - f_i(q(k+l)) \quad (1 \leq i \leq m) \quad (5.29)$$

とおくと、式 (5.7) の誤差評価関数は、

$$E(w, k) = \frac{1}{2} \sum_{l=0}^{N_B-1} \sum_{i=1}^m F_i(w, k, l)^2 \quad (5.30)$$

と表わせる。学習は、

$$F_i(w^*, k, l) = 0 \quad (1 \leq i \leq m, \forall k \in N_B Z, 0 \leq l \leq N_B - 1) \quad (5.31)$$

となるような最適点 w^* の探索である。ただし、

$$N_B Z = \{k = N_B \times r \mid r \in Z\}$$

とする。

²²⁾ 一般には、変数 q のスカラー値関数 $y(q(k))$ でもよい。

そこで、式(5.6)を

$$\mathbf{w}(k + N_B) = g(\mathbf{w}(k), k) = (g_1(\mathbf{w}(k), k), \dots, g_n(\mathbf{w}(k), k))^T \quad (5.32)$$

と表わし、 \mathbf{w}^* の近傍での線形近似系、

$$\mathbf{w}(k + N_B) = Dg(\mathbf{w}^*, k)\mathbf{w}(k) \quad (5.33)$$

の零解 $\mathbf{w}(k) = \mathbf{0}$ の安定性を考える。最適点 \mathbf{w}^* におけるヤコビ行列 Dg は、

$$Dg(\mathbf{w}^*, k) = \frac{1}{N_B m} \sum_{l=0}^{N_B-1} \sum_{i=1}^m A^{(i)}(k, l), \quad (5.34)$$

$$A^{(i)}(k, l) = I - N_B m \eta \mathbf{a}^{(i)}(k, l) \cdot \mathbf{a}^{(i)}(k, l)^T \quad (5.35)$$

と表わせる。ここで、

$$\mathbf{a}^{(i)}(k, l) = (a_{k,j,1}^{(i)}, \dots, a_{k,j,n}^{(i)})^T, \quad (5.36)$$

$$a_{k,j,i}^{(i)} = \frac{\partial F_i(\mathbf{w}^*, k, l)}{\partial w_j} \quad (1 \leq i \leq m, 1 \leq j \leq n, k \in N_B \mathbb{Z}, 0 \leq l \leq N_B - 1) \quad (5.37)$$

である。また、行列 $A^{(i)}(k, l)$ は固有値

$$\lambda^{(i)}(k, l) = 1 - N_B m \eta \|\mathbf{a}^{(i)}(k, l)\|^2 \quad (5.38)$$

の固有ベクトル $\mathbf{a}^{(i)}(k, l)$ と、これに直交する固有値1の $(n-1)$ 次元の固有空間 $V_k^{(i)}$ を持つことがBP法の場合と同様の考察によりわかる。

そこで、BP法の定理5-1に相当する次の定理が成立する。

[定理5-6] 時系列 $q(k)$ が周期 M を持ち、

$$\mathbf{a}^{(i)}(k, l) \neq \mathbf{0} \quad (1 \leq i \leq m, \forall k \in N_B \mathbb{Z}, 0 \leq l \leq N_B - 1 \text{ s.t. } 0 \leq k+l \leq M-1)$$

であるとし、

$$|\lambda^{(i)}(k, l)| = |1 - N_B m \eta \|\mathbf{a}^{(i)}(k, l)\|^2| < 1$$

$$(1 \leq i \leq m, \forall k \in N_B \mathbb{Z}, 0 \leq l \leq N_B - 1 \text{ s.t. } 0 \leq k+l \leq M-1)$$

となるよう η を十分小さく選んでおくとする。このとき、 $LN_B \geq M$ となる整数 L に対し、

$$B = Dg(\mathbf{w}^*, (L-1)N_B) \cdots Dg(\mathbf{w}^*, 0)$$

とすると、

(i) $mM < n$ の場合

B は固有値 1 を持ち、そのノルムは $\|B\| = 1$ である。

(ii) $mM \geq n$ の場合

$n \times mL N_B$ 行列 A を、

$$A = (a^{(1)}(0,0), \dots, a^{(m)}(0, N_B - 1), \dots, a^{(1)}((L-1)N_B, 0), \dots, a^{(m)}((L-1)N_B, N_B - 1))$$

と定義すると、

$$\|B\| < 1 \iff \text{rank} A = n.$$

(証明) $L N_B \geq M$ であるから、行列 A を構成する $m L N_B$ 本のベクトル $a^{(i)}(k, l)$ のうち線形独立なものは最大 mM 本であることに注意すれば、定理 5-1 と全く同様に証明できる。 \square

この定理より、関数誤差法による周期軌道の学習に関して、B P 法の場合の定理 5-2 に相当する次の定理が証明できる。

[定理 5-7] 離散力学系 (5.32) の不動点 w^* は定理 5-6 の条件および次の二つの条件を満たすとき一様漸近安定になる。

(条件 1) 教師信号 $q(k)$ は k に関して周期 M を持ち、 $mM \geq n$ である。

(条件 2) $\text{rank} A = n$ である。ここで、 A は定理 5-6 で定義した $n \times mL N_B$ 行列である。

(証明) まず、 $L_1 L N_B = L_2 M$ とする正整数 L_1, L_2 を選び、

$$B' = Dg(w^*, (L_1 L - 1) N_B) \cdots Dg(w^*, 0)$$

とする。このとき、定理 5-6 と条件 1、2 より、

$$\alpha = \|B'\| < 1$$

となる実数 α が存在することがわかる。この事実より定理 5-2 の証明と同様にして線形系 (5.33) の零解の一様漸近安定性が証明でき、補題 5-1 より (5.32) の不動点 w^* の一様漸近安定性が証明できる。 \square

定理 5-7 の条件 1 は N_B に依存しない。また、条件 1 が成立していれば条件 2 は一般に成立する。したがって、関数誤差法による周期軌道の学習は、 N_B に依存せず、

学習対象の次元と軌道の周期の積がパラメータの総数より多い場合は一般に局所収束することを定理5-7は意味している。

また、定理5-6、5-7の証明において重要なのは、BP法の場合と同じく学習対象の次元と教師信号の周期、パラメータの総数の間の関係であり、 N_B の値は本質的ではない。このことから、BP法の場合の定理5-3~5-5は関数誤差法に関してもそのまま成立することがわかる。

さらに、学習係数 η のBP法で N_B 回学習すると、最適点 w^* の近傍で、

$$\begin{aligned} Dg(w^*, k + N_B - 1) \cdots Dg(w^*, k) &= \prod_{i=0}^{N_B-1} (I - \eta \sum_{l=0}^m a^{(i)}(k+l) a^{(i)}(k+l)^T) \\ &= \prod_{i=0}^{N_B-1} (I - \eta \sum_{l=0}^m a^{(i)}(k, l) a^{(i)}(k, l)^T) \\ &= \frac{1}{N_B m} \sum_{l=0}^{N_B-1} \sum_{i=1}^m A^{(i)}(k, l) + O(\eta^2) \end{aligned} \quad (5.39)$$

となって、式(5.34)で表される関数誤差法による1回当たりの学習のヤコビ行列に η^2 のオーダーを無視すれば一致している。したがって、第3章において確率的降下法の理論により示された、学習係数 η の関数誤差法と学習係数 $N_B \cdot \eta$ のBP法との等価性は学習の最終状態においても成立していることがわかる。

軌道誤差法による学習

軌道誤差法による学習則は式(5.6)と(5.8)で定義される。また、

$$F_i(w(k), k, l) = f_b^{i+1}(q(k), w(k)) - f_b^{i+1}(q(k)) \quad (1 \leq i \leq m) \quad (5.40)$$

とおくと、式(5.8)の誤差評価関数は、

$$E(w, k) = \frac{1}{2} \sum_{l=0}^{N_B-1} \sum_{i=1}^m F_i(w, k, l)^2 \quad (5.41)$$

と表わせ、関数誤差法と同じ形式になる。学習は、

$$F_i(w^*, k, l) = 0 \quad (1 \leq i \leq m, \forall k \in N_B \mathbb{Z}, 0 \leq l \leq N_B - 1) \quad (5.42)$$

となるような最適点 w^* の探索である。

そこで、式(5.6)を関数誤差法と同様に式(5.32)で表わすと、最適点 w^* におけるそのヤコビ行列 Dg は式(5.34)および(5.35)で表わせる。

以上のように、軌道誤差法による学習は関数誤差法の場合と全く同様に定式化できる。軌道誤差法が関数誤差法と異なるのは、教師信号 $q(k)$ が周期的であっても、 $F_i(w^*, k, l)$ は k に関して周期的とは限らない点である。このため、BP法の場合の定理5-1と関数誤差法の場合の定理5-6は、軌道誤差法の場合は次のかたちで成立する。

[定理5-8]

$$a^{(i)}(k, l) \neq 0 \quad (1 \leq i \leq m, \forall k \in N_B \mathbb{Z}, 0 \leq l \leq N_B - 1)$$

であるとし、

$$|\lambda^{(i)}(k, l)| = |1 - N_B m \eta| |a^{(i)}(k, l)|^2 < 1 \quad (1 \leq i \leq m, \forall k \in N_B \mathbb{Z}, 0 \leq l \leq N_B - 1)$$

となるよう η を十分小さく選んでおくとする。このとき、任意の整数 k および正整数 L に対し、

$$B(k) = Dg(w^*, k + (L-1)N_B) \cdots Dg(w^*, k)$$

とすると、

(i) $mLN_B < n$ の場合

$B(k)$ は固有値1を持ち、そのノルムは $\|B(k)\| = 1$ である。

(ii) $mLN_B \geq n$ の場合

$n \times mLN_B$ 行列 $A(k)$ を、

$$A(k) = (a^{(1)}(k, 0), \dots, a^{(m)}(k, N_B - 1), \dots, a^{(1)}(k + (L-1)N_B, 0), \dots, a^{(m)}(k + (L-1)N_B, N_B - 1))$$

と定義すると、

$$\|B(k)\| < 1 \Leftrightarrow \text{rank} A(k) = n.$$

□

証明は定理5-1と同様である。

特に教師信号 $q(k)$ が周期 M を持つとき、定理5-8の行列 $B(k)$ も M 周期的であるから、 LN_B が M の整数倍になるような L に対して定理5-8を適用すれば、軌道誤差法による周期軌道の学習に関しても定理5-7が成立することがわかる(ただし、定理5-7の行列 A は定理5-8の $A(0)$ とする)。

同様に、定理5-3、定理5-5が軌道誤差法による学習に関してもそのまま成立する。しかし、定理5-4は一般には成立しない。これは、教師信号が周期的であっても

$F_i(\mathbf{w}^*, k, l)$ は k に関して周期的とは限らないため、カオスの中の不安定周期軌道を用いた議論ができないからである。

5.3 連続力学系の学習の収束性

5.3.1 学習を表わす差分方程式

m 次元連続力学系、

$$\frac{dq(t)}{dt} = f(q(t)) \quad (q \in \mathbb{R}^m) \quad (5.43)$$

を次式で記述される RNN で学習する問題を考える。

$$\frac{dx(t)}{dt} = \hat{f}(x(t), \mathbf{w}) \quad (x \in \mathbb{R}^m, \mathbf{w} \in \mathbb{R}^n). \quad (5.44)$$

ここで、関数 f と \hat{f} はそれぞれ、 q と (x, \mathbf{w}) に関する滑らかな関数とする。 \mathbf{w} は RNN のパラメータを表わす n 次元ベクトルである。

学習アルゴリズムとして、ここでは第3章で定義した最適制御型学習則の速度誤差法と波形誤差法を考える。

最適制御型学習則は、誤差を評価する一定周期 T_B を設け、教師信号と RNN の出力との誤差を1周期積分したものを評価関数 E とし、これを減少させるようパラメータを修正していく方法であった。パラメータの更新則は T_B の整数倍の時間だけ値を持つ次のような差分方程式（離散力学系）になる。

$$\mathbf{w}((k+1)T_B) = \mathbf{w}(kT_B) - \eta \nabla_{\mathbf{w}} E(\mathbf{w}(kT_B), k). \quad (5.45)$$

ここで、 η は学習係数である。速度誤差法、関数誤差法の誤差評価関数はそれぞれ次式のように定義される。

(1) 速度誤差法

$$E(\mathbf{w}, k) = \frac{1}{2} \int_{kT_B}^{(k+1)T_B} \|\hat{f}(q(t), \mathbf{w}) - f(q(t))\|^2 dt. \quad (5.46)$$

(2) 波形誤差法

$$E(\mathbf{w}, k) = \frac{1}{2} \int_{kT_B}^{(k+1)T_B} \|\mathbf{x}(t) - q(t)\|^2 dt, \quad (5.47)$$

$$\frac{d\mathbf{x}(t)}{dt} = \hat{f}(\mathbf{x}(t), \mathbf{w}), \quad \mathbf{x}(kT_B) = q(kT_B). \quad (5.48)$$

5.3.2 最適制御型学習則の収束条件

速度誤差法の収束条件

速度誤差法による学習則は、式(5.45)、(5.46)より次のように表現できる。

$$\mathbf{w}(k+1) = \mathbf{w}(k) - \frac{\eta}{2} \sum_{i=1}^m \int_{kT_B}^{(k+1)T_B} \nabla_{\mathbf{w}} F_i(\mathbf{w}(k), \mathbf{q}(t))^2 dt. \quad (5.49)$$

ここで、

$$F_i(\mathbf{w}(k), \mathbf{q}(t)) = \hat{f}_i(\mathbf{q}(t), \mathbf{w}) - f_i(\mathbf{q}(t)) \quad (1 \leq i \leq m) \quad (5.50)$$

であり、 $\mathbf{w}(kT_B)$ を $\mathbf{w}(k)$ と略記している。学習は、

$$F_i(\mathbf{w}^*, \mathbf{q}(t)) = 0 \quad (\forall t \in \mathbf{R}, 1 \leq i \leq m) \quad (5.51)$$

となるような最適点 \mathbf{w}^* を勾配降下法によって探索することである。 \mathbf{w}^* は差分方程式(5.49)の不動点であるから、その安定性が学習の局所収束性を意味することになる。

そこで、以下、最適点 \mathbf{w}^* が少なくとも一つ存在するものとして、速度誤差法による学習の \mathbf{w}^* への局所収束性を差分方程式(5.49)の不動点としてのその安定性と考察解析する。

安定性解析には離散力学系の学習と同様に線形近似解析を用いる。まず、不動点の近傍で(5.49)を次式のように線形化する。

$$\mathbf{w}(k+1) = B(k)\mathbf{w}(k). \quad (5.52)$$

ここで、

$$B(k) = \frac{1}{mT_B} \sum_{i=1}^m \int_{kT_B}^{(k+1)T_B} B^{(i)}(t) dt, \quad (5.53)$$

$$B^{(i)}(t) = I - \eta m T_B \mathbf{a}^{(i)}(t) \mathbf{a}^{(i)}(t)^T, \quad (5.54)$$

$$\mathbf{a}^{(i)}(t) = (a_1^{(i)}(t), \dots, a_n^{(i)}(t))^T, \quad a_j^{(i)}(t) = \frac{\partial F_i}{\partial w_j}(\mathbf{w}^*, \mathbf{q}(t)) \quad (1 \leq i \leq m, 1 \leq j \leq n)$$

である。さらに、

$$\lambda^{(i)}(t) = 1 - \eta m T_B \|\mathbf{a}^{(i)}(t)\|^2 \quad (1 \leq i \leq m) \quad (5.55)$$

とすると、各 $B^{(i)}(t)$ は固有値 $\lambda^{(i)}(t)$ の固有ベクトル $\mathbf{a}^{(i)}(t)$ と、これに直交する、固有値 1 の $(n-1)$ 次元固有空間 $V^{(i)}(t)$ を持つ。このことは離散系学習の場合の行列 $A^{(i)}(k)$ の性質と同様に証明できる。

また、教師信号は一般に有界なので、式(5.55)より、学習係数 η を十分小さく選んでおけば、任意の i と j について、

$$|\lambda^{(i)}(t)| < 1 \quad (5.56)$$

となる。 $B^{(i)}(t)$ に関する上述の性質より、このとき

$$\|B^{(i)}(t)\| \leq 1 \quad (1 \leq i \leq m, \forall t \in \mathbb{R})$$

である。したがって式(5.52)、(5.53)より、

$$\begin{aligned} \|w(k+1)\| &= \|B(k)w(k)\| \\ &\leq \frac{1}{mT_B} \sum_{i=1}^m \int_{kT_B}^{(k+1)T_B} \|B^{(i)}(t)w(k)\| dt \leq \|w(k)\| \end{aligned}$$

である。よって、 $|\lambda^{(i)}(t)| < 1$ であれば、

$$\|B(k)\| \leq 1 \quad (5.57)$$

である。このとき、線形近似系(5.52)の零点に関して、補題5-1の下に記した一様漸近安定性の条件のうち、(i)の一様安定性は満足していることがわかる。

そこで、条件(5.56)が成立しているという仮定のもとに、線形近似系の零点が一様漸近安定性の条件(ii)を満足するかどうかを検討する。具体的には、

$$B = B(k+L-1) \cdot B(k+L-2) \cdots B(k)$$

で定義される行列 B のノルムが

$$\|B\| < \beta$$

を満足するような、離散時間 k に依存しない整数 L と $0 < \beta < 1$ となる実数 β が存在するかどうかをチェックする。このような L と β があれば、不等式(5.57)と式(5.52)の線形性より、条件(ii)の成立が示せるからである。

以下、教師信号が(1)平衡点、(2)周期軌道、(3)概周期軌道、(4)カオスの各場合に分けて考察を進める。

(1) 平衡点の学習

この場合は $B^{(i)}(t)$ は時間 t に依存しないので、式(5.53)の $B(k)$ は、

$$B(k) = \frac{1}{m} \sum_{i=1}^m [I - \eta m T_B a^{(i)}(0) a^{(i)}(0)^T] \quad (5.58)$$

となって、 k に依存しない。したがって、

(i) $m < n$ の場合

$V^{(1)}(0) \cap \dots \cap V^{(m)}(0)$ は非零ベクトルを含むから、 $\|B(k)\| \geq 1$ となって、線形近似系 (5.52) の零点は漸近安定ではない。

(ii) $m \geq n$ の場合

$a^{(1)}(0), \dots, a^{(m)}(0)$ は一般に n 本の一次独立なベクトルを含むから、 $V^{(1)}(0) \cap \dots \cap V^{(m)}(0) = \{0\}$ となる。よって、 $|\lambda^{(i)}(t)| < 1$ となるように学習係数 η を十分小さく選んでおけば、 $\|B(k)\| < 1$ となって (5.52) の零点は一様漸近安定になる。以上と補題 5-1 より、次の定理が成り立つ。

[定理 5-0] 平衡点を学習する場合 (関数 $F_i(w, q(t))$ が時間 t に依存しない場合)、すべての i と t について $|\lambda^{(i)}(t)| < 1$ であれば、 $m \geq n$ の場合、学習を表わす差分方程式 (5.49) の不動点は一般に一様漸近安定になる。すなわち、平衡点の学習は最適値に局所収束する。 $m < n$ の場合は線形近似系の零点は漸近安定ではない。 □

(2) 周期軌道の学習

まず、教師信号 $q(t)$ が時間的に変動している場合に一般に成立する補題を二つ述べておく。

[補題 5-2] すべての i と t について $|\lambda^{(i)}(t)| < 1$ とする。このとき、 $\gamma > 0$ とし、時間区間 $[t, t + \gamma]$ で $a^{(i)}(t)$ が一定値でない場合、 $0 < \beta < 1$ となる β が一般に存在し、

$$\hat{B}^{(i)} = \frac{1}{\gamma} \int_t^{t+\gamma} B^{(i)}(s) ds$$

と表わすと、

$$\|\hat{B}^{(i)}\| < \beta$$

である。

(証明) 仮定と $q(t)$ の連続性より、

$$t < t_1 < t_2 < \dots < t_p < t + \gamma$$

となる p 個 ($p > n$) の点 t_j を、 $a^{(i)}(t_j)$ の値が互いに相異なるように選べる。このとき一般に

$$V^{(i)}(t_1) \cap \dots \cap V^{(i)}(t_p) = \{0\}$$

となる。したがって、

$$S^{n-1} = \{ v \in \mathbb{R}^n \mid \|v\| = 1 \}$$

とすると、任意の $v \in S^{n-1}$ に対して、 t_j があって、 $\|B^{(i)}(t_j)v\| < \|v\|$ となる。ここで、 t_j は $[t, t+\gamma]$ の内点で、また $B^{(i)}(t)$ は t について連続、さらに $B^{(i)}(t)v$ は v について連続であるから、 $\epsilon > 0$ 、 $0 < \alpha < 1$ 、 $\delta > 0$ となる実数 ϵ 、 α 、 δ が存在して、 $\hat{U}(t_j; \epsilon) \subset [t, t+\gamma]$ で、任意の $t' \in \hat{U}(t_j; \epsilon)$ 、 $v' \in U(v; \delta)$ に対して、

$$\|B^{(i)}(t')v'\| < \alpha\|v'\|$$

となる。ここで、 $\hat{U}(t_j; \epsilon)$ と $U(v; \delta)$ は、 \mathbb{R} における t_j の ϵ -開近傍、 \mathbb{R}^n における v の δ -開近傍をそれぞれ表す。よって、

$$\begin{aligned} \|\hat{B}^{(i)}v'\| &\leq \frac{1}{\gamma} \left[\int_{t_j-\epsilon}^{t_j+\epsilon} \|B^{(i)}(s)v'\| ds + \int_{[t, t+\gamma] - \hat{U}(t_j; \epsilon)} \|B^{(i)}(s)v'\| ds \right] \\ &\leq \frac{1}{\gamma} (2\epsilon\alpha + \gamma - 2\epsilon) = 1 - \frac{2\epsilon(1-\alpha)}{\gamma} < 1 \end{aligned}$$

ここで、 ϵ 、 α 、 δ は v に依存するので、

$$\beta(v) = \frac{1}{\gamma} (2\epsilon\alpha + \gamma - 2\epsilon), \quad \delta = \delta(v)$$

と表す。今、 $v \in S^{n-1}$ は任意であるから各 $v \in S^{n-1}$ に対してこのような $\beta(v)$ と $\delta(v)$ がとれる。また、 S^{n-1} はコンパクトだから、有限個の $v_1, \dots, v_q \in S^{n-1}$ があって、

$$\bigcup_{j=1}^q U(v_j; \delta(v_j)) \supset S^{n-1}$$

となる。そこで、

$$\beta = \max\{ \beta(v_j) \mid 1 \leq j \leq q \}$$

とすると、 $\beta < 1$ であり、任意の $v \in S^{n-1}$ について、 $\|B^{(i)}(t)v\| < \beta\|v\|$ となる。よって、

$$\|\hat{B}^{(i)}\| < \beta$$

となる。 □

[補題5-3] 任意の正の実数 τ と T について、

$$S = R/\tau \quad (\text{区間 } [0, \tau] \text{ の両端を同一視した円})$$

とし、 $0 < t_1 < t_2 < \tau$ かつ $t_2 - t_1 < T$ となる t_1 と t_2 について

$$J = [t_1, t_2] \subset S$$

とすると、正整数 l があり、任意の整数 k について $0 \leq j \leq l-1$ となる j が存在して、

$$[(k+j)T, (k+j+1)T] \supset J$$

となる。

(証明)

$$T' = \frac{1}{2} \min(t_1, T - t_2 + t_1)$$

とし、 $t_3 = t_1 - T'$ とする。

このとき、

$$G: S \rightarrow S$$

$$t \mapsto t + T \pmod{\tau}$$

は概周期的になるから、整数の集合 Z_1 と正整数の有限集合 Z_2 があって、任意の $k \in Z_1$ について $G^k(0) \in U(t_3; T')$ であり、かつ、

$$Z = Z_1 + Z_2 = \{ z_1 + z_2 \mid z_1 \in Z_1, z_2 \in Z_2 \}$$

となる¹⁵⁾。そこで、 $l = \max Z_2$ とすれば l は題意を満たす。 \square

この二つの補題より、周期軌道の学習が一般に収束することが証明できる。

[定理 5-10] 関数 $F_t(w, q(t))$ が時間 t に関して (最小) 周期 $\tau (> 0)$ を持つならば、 $|\lambda^{(0)}(t)| < 1$ であれば、差分方程式 (5.49) の不動点 w^* は一般に一様漸近安定である。すなわち、周期軌道の学習は一般に最適値に局所収束する。

(証明) 教師信号 $q(t)$ の周期 τ と学習の周期 T_B の比 τ/T_B が有理数の場合と無理数の場合に分けて、線形近似系 (5.52) の零点の一様漸近安定性を証明する。これを証明すれば、 w^* の一様漸近安定性も補題 5-1 より保証される。

(i) τ/T_B が有理数の場合

次式をみたす整数 L_T と L_r が存在する。

$$T_B \cdot L_T = \tau \cdot L_r.$$

ここで、

$$B = B(k + L_T - 1) \cdot B(k + L_T - 2) \cdots B(k) \quad (5.59)$$

とすると、 $q(t)$ の周期性より B は k に依存しない。よって、線形系 (5.52) の解は、

$$w(k + L_T) = Bw(k)$$

をみす。さらに、(5.57) より、 $\|B(k)\| \leq 1$ であるから、 $\|B\| < 1$ であれば (5.52) の零点は一様漸近安定になる。

ここで、 $q(t)$ は周期 τ だから、区間 $[lT_B, (l+1)T_B]$ において $q(t)$ が一定とならないような l ($0 \leq l \leq L_T - 1$) がある。この区間では $q(t)$ が一定ではないから、ある i があって、 $a^{(i)}(t)$ も時間変化する。よって補題 5-2 より $0 < \beta < 1$ となる β が存在して、

$$\frac{1}{T_B} \left\| \int_{lT_B}^{(l+1)T_B} B^{(i)}(t) dt \right\| < \beta$$

となる。したがって、任意の $v \in \mathbb{R}^n$ について、

$$\|Bv\| < \left(1 - \frac{1-\beta}{m}\right) \|v\| < \|v\|$$

なので、 $\|B\| < 1$ である。よって (5.52) の零点は一様漸近安定である。

(ii) τ/T_B が無理数の場合

教師信号 $q(t)$ は周期 τ の周期関数なので、 $0 < t_1 < t_2 < \tau$ かつ $t_2 - t_1 < T_B$ となる t_1 と t_2 で、

$$J = [t_1, t_2] \subset S$$

上で $q(t)$ が一定とならないようなものが存在する。このとき、補題 5-3 より、正整数 l があり、任意の整数 k について $0 \leq j \leq l-1$ となる j が存在して、

$$[(k+j)T_B, (k+j+1)T_B] \cap J$$

となる。また、補題 5-2 より $0 < \beta < 1$ となる β が存在して、

$$\frac{1}{t_2 - t_1} \left\| \int_{t_1}^{t_2} B^{(i)}(t) dt \right\| < \beta.$$

したがって、

$$B = B(k+l-1) \cdot B(k+l-2) \cdots B(k)$$

とすると、 $0 < \alpha < 1$ であつ k に依存しない α が存在して、

$$\|B\| \leq \alpha$$

となる。よって (5.52) の零点は一様漸近安定である。

□

(3) 概周期軌道の学習

概周期軌道の学習も一般に収束する。周期軌道はもちろん概周期軌道であるから、次の定理は定理 5-10 の拡張である。

[定理 5-11] 関数 $F_i(w, q(t))$ が時間 t に関して概周期的とする。このとき、任意の i, t について $|\lambda^{(i)}(t)| < 1$ であれば、差分方程式 (5.49) の不動点 w^* は一般に一様漸近安定である。すなわち、概周期軌道の学習は一般に最適値に局所収束する。

(証明) $q(t)$ は概周期的であるから、区間 $[\sigma, \sigma + \gamma]$ で $q(t)$ が一定とならないような実数 σ および $\gamma (> 0)$ が存在する。補題 5-2 より、ある $i (1 \leq i \leq m)$ に対して $0 < \beta < 1$ となる β が存在して、

$$\frac{1}{\gamma} \left\| \int_{\sigma}^{\sigma+\gamma} B^{(i)}(t) dt \right\| < \beta$$

となる。 γ はいくらでも小さくできるから、 $\gamma < T_B$ としておく。

$B^{(i)}(t)$ は $q(t)$ に関して連続だから、正数 ϵ があって、 $\|q(\sigma') - q(\sigma)\| < \epsilon$ ならば、

$$\frac{1}{\gamma} \left\| \int_{\sigma'}^{\sigma'+\gamma} B^{(i)}(t) dt - \int_{\sigma}^{\sigma+\gamma} B^{(i)}(t) dt \right\| < \frac{1-\beta}{3}$$

となる。このとき、

$$\frac{1}{\gamma} \left\| \int_{\sigma'}^{\sigma'+\gamma} B^{(i)}(t) dt \right\| \leq \beta' < 1$$

である。ただし、 $\beta' = (2\beta + 1)/3$ である。

ここで、連続力学系 (1) の time- T_B -map、

$$H : x(t) \mapsto x(t + T_B) \quad (5.60)$$

は、教師信号 $q(t)$ が概周期的だから、

$$A_T = \overline{\{q(t)\} - \infty < t < +\infty}$$

上では、やはり概周期的になる。よって、正整数 l があり、任意の整数 k について $0 \leq j \leq l-1$ となる j が存在して、

$$\|q((k+j)T_B) - q(\sigma)\| < \epsilon$$

となる。したがって、周期軌道の場合と同様に、

$$B = B(k+l-1) \cdot B(k+l-2) \cdots B(k)$$

とすると、

$$\|B\| \leq 1 - \frac{\gamma(1-\beta^j)}{mT_B} < 1$$

となる。したがって(5.52)の零点は一樣漸近安定であり、補題5-1より、 w^* も一樣漸近安定である。□

(4) カオスの学習

離散力学系のカオスの学習に関しては、線形近似系が一樣漸近安定性を持たないため、収束の証明は成功していない。連続系のカオスについても同様の困難があり、線形近似系はやはり一般には一樣漸近安定性を持たない。

カオス学習の収束性の検討に入る前に、ここで、カオス軌道の持つ性質をまとめておく。連続系のカオスも前節の離散系と同じ定義を用いる。すなわち、 $q(t)$ をカオス軌道とし、

$$A_T = \overline{\{q(t) \mid -\infty < t < +\infty\}}$$

とすると、(a) A_T はコンパクト、(b) $q(t)$ は非周期的、(c) A_T は不安定な周期軌道を含む、とする。

まず次の定理が成り立つ。この定理は離散系のカオスの学習において線形近似系の一樣漸近安定性が成り立たないことを示す定理5-4と本質的に同じである。

[定理5-12] 集合 A_T が平衡点 \bar{q}_0 を含み、 $q(t)$ は A_T で稠密な軌道とする。さらに、 $n > m$ とする。このとき、線形近似系(5.52)の零点は一樣漸近安定ではない。

(証明) 定理5-9で示したように、 $q(t) \equiv \bar{q}_0$ の場合は $\|B(k)\| = 1$ になる。また、 $B^{(l)}(t)$ の q に関する連続性および行列のノルムの連続性より、任意の実数 T_B と任意の正整数 L に対し正数 β が存在して、任意の整数 k_0 について、

$$\|q(k_0 T_B) - \bar{q}_0\| < \beta$$

ならば、

$$\|B(k_0 + L - 1) \cdot B(k_0 + L - 2) \cdots B(k_0)\| > \frac{1}{2}$$

となる。 $q(t)$ は A_T で稠密だから、このような k_0 は必ず存在する。したがって、任意の正数 δ に対し、 $\epsilon = (1/4)\delta$ とすると、任意の L に対し、上記のような k_0 と $(1/2)\delta < \|w_0\| < \delta$ となる w_0 を選ぶと、

$$\|B(k_0 + L - 1) \cdot B(k_0 + L - 2) \cdots B(k_0) \cdot w_0\| > \epsilon$$

となる。よって、(5.52)の零解は一様漸近安定ではない。 □

しかし、線形近似系(5.52)の零解が漸近安定であることは、以下のように示せる。

[定理 5-13] 式(5.60)の time- T_B -map H が A_T 上の不変確率測度 P に関してエルゴード的で、

$$P(\{q(0) \mid \|B(0)\| < 1\}) > 0 \quad (5.61)$$

であるとする。このとき、任意の i, l について $|\lambda^{(i)}(t)| < 1$ であれば、(5.52)の零解が漸近安定であることが $q(t)$ の初期値に関して確率 1 で成り立つ。 □

証明は離散系の場合の定理 5-5 と同様である。

定理 5-5 の条件のうち、 H のエルゴード性は $q(t)$ が混合的であれば成立する。また条件 (5.61) は、時間区間 $[0, T_B]$ で $q(t)$ が一定でない確率が零でなければ一般に成立する。したがって、(5.52)の零解は、一般的に漸近安定であることがわかる。

また、この定理の系として次が成立する。

[系] 式(5.49)において、関数 F_i がパラメータ w に関して線形であり、かつ定理 5-13 の条件が成り立てば、 w^* は $q(t)$ の初期値に関して確率 1 で漸近安定になる。 □

定理 5-12 で本質的なのは、前述のカオスの性質 (c) である。すなわち、アトラクタが不安定な周期軌道を含むことにより、稠密な軌道はこの周期軌道の任意の近傍をとおる。このため、収束の速度が $q(t)$ の初期値に対して一様でなく、一様漸近安定性が保証されない。もちろん、 $q(t)$ は一般にエルゴード的であるから、十分時間がたてば、 $q(t)$ は教師信号としてアトラクタ全体の情報を反映する。これが定理 5-13 の意味するところである。重要なのは、エルゴード性は時間平均と空間平均の一致を保証するが、時間平均の収束の速さの初期値に対する一様性は保証しない点である。

波形誤差法の収束条件

波形誤差法による学習則は、(5.45)、(5.47) より次のように表現できる。

$$w(k+1) = w(k) - \frac{\eta}{2} \sum_{i=1}^m \int_{kT_B}^{(k+1)T_B} \nabla_w F_i(w(k), q(t), t)^2 dt. \quad (5.62)$$

ここで、

$$F_i(w(k), q(t), t) = x_i(t, w(k)) - q_i(t) \quad (1 \leq i \leq m) \quad (5.63)$$

であり、 $x(t, w(k))$ は $w = w(k)$ とした微分方程式 (5.48) の解である。

学習の局所収束性は、速度誤差法と同様、

$$F_i(w^*, q(t), t) = 0 \quad (\forall t \in \mathbb{R}, 1 \leq i \leq m) \quad (5.64)$$

となるような、差分方程式 (5.62) の不動点 w^* の安定性として解析される。

波形誤差法では、 $x(t, w(k))$ と $q(t)$ との誤差を評価しているから、速度誤差法とは違って F_i の時間成分は $q(t)$ だけでは定まらない。 F_i の最後の引数 t は $x(t, w(k))$ に依存する成分を表わしている。この点のみ考慮すれば、不動点近傍での線形化方程式に関して、速度誤差法の場合の結果がほぼすべて成立する。すなわち、(5.62) の線形化方程式は (5.52) で表わされ、(5.53) ~ (5.55) の関係式は、 $B^{(i)}(t)$ 、 $a^{(i)}(t)$ 、 $\lambda^{(i)}(t)$ 、 $V^{(i)}(t)$ を $B^{(i)}(q(t), t)$ 、 $a^{(i)}(q(t), t)$ 、 $\lambda^{(i)}(q(t), t)$ 、 $V^{(i)}(q(t), t)$ にそれぞれ置き換えればすべて成立する。また、すべての i と t について $|\lambda^{(i)}(q(t), t)| < 1$ であれば、(5.57) も成り立つ。

したがって、定理 5-9、定理 5-11、定理 5-13 が、波形誤差法による学習についても、それぞれ、次の定理 5-14、定理 5-15、定理 5-16 のかたちで成立する。周期軌道は概周期軌道でもあるので、その収束性は定理 5-15 によって保証される。よって、定理 5-10 に相当する定理については省略する。定理 5-12 は波形誤差法の場合は成立しない。これは、平衡点の近傍を学習している場合、つまり $q(t)$ がほぼ一定値の場合であっても $B^{(i)}(q(t), t)$ は一般には時間変化するため $\|B(k)\| \approx 1$ とは限らないからである。

[定理 5-14] 平衡点を波形誤差法で学習する場合 (関数 $F_i(w, q(t), t)$ の $q(t)$ が時間 t に依存しない場合)、すべての i と $t \in [0, T_B]$ について $|\lambda^{(i)}(q(t), t)| < 1$ であれば、

- (a) $B^{(i)}(q(t), t)$ が $[0, T_B]$ で一定ならば、 $m \geq n$ の場合、学習を表わす差分方程式 (5.62) の不動点は一様漸近安定になる。すなわち、平衡点の学習は最適値に局所収束する。 $m < n$ の場合は線形近似系の零点は漸近安定ではない。
- (b) $B^{(i)}(q(t), t)$ が $[0, T_B]$ で一定でなければ、差分方程式 (5.62) の不動点は一様漸近安定になる。 □

(a) の証明は [定理 5-9] と同じである。(b) は補題 5-2 より、定理 5-11 と同様に証明

できる。

[定理5-15] $q(t)$ が時間 t に関して概周期的とする。このとき任意の $i, t \in [0, T_B]$ について $|\lambda^{(i)}(q(t), t)| < 1$ ならば、差分方程式(5.62)の不動点 w^* は一般に一様漸近安定である。すなわち、波形誤差法による概周期軌道の学習は一般に最適値に局所収束する。□

$B^{(i)}(q(t), t)$ の時間変化は $q(t)$ だけでは定まらないが、 $q(kT_B)$ に関して $B^{(i)}(q(t), t)$ は連続なので、定理3と同様の証明が行なえる。

[定理5-16] 式(5.60)の time- T_B -map H が A_T 上の不変確率測度 P に関してエルゴード的で、

$$P(\{q(0) \mid \|B(0)\| < 1\}) > 0 \quad (5.65)$$

であるとする。このとき、任意の $i, q \in A_T, t \in [0, T_B]$ について $|\lambda^{(i)}(q(t), t)| < 1$ であれば、式(5.62)の線形化方程式(5.52)の零解が漸近安定であることが $q(t)$ の初期値に関して確率1で成り立つ。□

証明は定理5-13と同じである。 $|\lambda^{(i)}(q(t), t)| < 1$ の条件は、 $A_T \times [0, T_B]$ がコンパクトなので、 η を十分小さく選べば成立する。

5.4 結言

逐次型のBP法のような二乗誤差最小化の勾配降下法による力学系の学習をシステムパラメータを変数とする非自律的力学系の状態変化としてとらえ、最適点近傍での学習の収束条件を不動点の安定性として解析的に考察した。

本章で得られた結果は表5-1のように要約される。

まず、離散力学系の逐次型BP法による学習の最適点への収束性に関して、以下のことを証明した。周期軌道の学習は、学習対象の次元と周期の積がパラメータ空間の次元より大きい場合には局所収束する。非周期的な概周期軌道の学習は一般に局所収束する。カオス軌道の学習は、線形近似系の零点の一様漸近安定性が一般には成立しないため、一般的な収束条件は明らかでないが、線形近似系の零点の漸近安定性は確率1で保証される。したがって、線形のRNNでカオスを学習する場合には、一般的な仮定のもとに学習は最適点に収束する。この結果は関数誤

表 5-1 離散系と連続系の学習の局所収束性のまとめ

	離散系 (B P法、関数誤差法、軌道誤差法)	連続系 (速度誤差法、波形誤差法)
平衡点	(周期軌道に含まれる)	$m \geq n$: 一般に収束 $m < n$: ? (定理 5-9、5-14)
周期軌道	$m \cdot M \geq n$: 一般に収束 $m \cdot M < n$: ? (M = 周期) (定理 5-2、5-7)	一般に収束 (定理 5-10)
概周期軌道	非周期的ならば収束 (定理 5-3)	一般に収束 (定理 5-11、5-15)
カオス	周期 n/m 未満の周期点を含むと 線形近似系は一樣漸近安定でない (定理 5-4 (軌道誤差法は×)) 線形近似系は一般に確率 1 で収束 (定理 5-5) 一般には ?	平衡点を含むと 線形近似系は一樣漸近安定でない (定理 5-12 (波形誤差法は×)) 線形近似系は一般に確率 1 で収束 (定理 5-13、5-16) 一般には ?

m : 教師信号の次元
 n : パラメータの個数

差法・軌道誤差法による学習にもほぼそのまま拡張できる。

また、連続力学系の最適制御型学習則による学習の局所収束性に関しては、以下のことを証明した。速度誤差法による学習では、学習対象のアトラクタが、周期軌道、概周期軌道の場合は一般に、平衡点の場合はパラメータの個数が学習対象の系の次元より小さければ収束が保証される。また、カオスの学習については、離散系の学習と同様、線形近似解析が困難であり、線形近似系の形な RNN による学習の局所収束性のみ保証される。この結果は波形誤差法による学習の場合も、ほぼそのまま成立する。

カオス時系列の学習が解析困難であるのは次の理由による。すなわち、学習アルゴリズムが非周期的な時間関数に駆動される非自律的差分方程式であるため、線形近似系の零点の一樣漸近安定性を示す必要がある。しかし、定理 5-3 の証明で

述べたように、カオス軌道がその閉包に無数の周期軌道を含むことが線形近似系の一様漸近安定性を成立させなくしているのである。このように、カオス学習の収束判定は線形近似を受けつけない、本質的に非線形な問題である。

カオスによって駆動される非自律系の不動点の安定条件は、カオスの同期条件^{46, 47, 48)}と全く同型になる。また、カオスアトラクタの含む不安定周期軌道がBP法によるカオス時系列の学習において重要な役割を演じているという興味深い報告もある³³⁾。このように、カオス時系列の学習はNNによるシステムの同定や決定論的予測などの応用的な問題にとどまらず、カオスの本質的な問題に直結している可能性がある。

第6章 学習収束過程と最適状態の引力圏

6.1 緒言

本章では、簡単なRNNモデルを用いて、1次元離散力学系の学習の収束過程を数値的に解析する。

前章で考察したように、逐次型勾配降下法による力学系の学習は結合荷重空間上の非自律的離散力学系と考えることができる。この力学系の不動点としての最適点の近傍での学習収束過程を数値実験により調べ、不動点の特性指数とリアプノフ指数との関係を考察する。

また、カオス時系列の学習は、カオス時系列によって励振された非自律的力学系であり、前章の定理で述べたように、周期励振系のみならず概周期励振系とも異なった性質を持つ。このため、さまざまな興味深い現象が観測されるものと考えられる。その一例として、カオス時系列の学習においては、リアプノフ指数が負であっても不動点がリアプノフの意味では安定にならない場合があることを指摘する。また、カオスの学習が不安定になった場合の結合荷重空間上のアトラクタについて考察する。

さらに、逐次型勾配降下法による力学系の学習を表現する簡単なモデルを構成し、その最適点が riddled basin⁴⁸⁾ と呼ばれる非常に複雑な引力圏を伴う不動点になっていることを明らかにする。

6.2 数値実験

6.2.1 離散力学系学習の収束過程

ここでは、次式で表わされるRNNを用いて、1次元離散力学系学習の収束過程を数値的に考察する。

$$\begin{aligned}x(k+1) &= \hat{f}(x(k), \omega), \\ \hat{f}(x(k), \omega) &= \sum_{m=1}^L W_m^A G(W_m^B x(k)), \\ \omega &= (W_1^A, \dots, W_L^A, W_1^B, \dots, W_L^B).\end{aligned}\tag{6.1}$$

このRNNは第2章で述べた、離散力学系の学習を対象としたものであり、 x はその出力変数、 W_m^a と W_m^b は結合荷重を表わすパラメータで、これらを成分として持つベクトルを w で表現する。 G はシグモイド関数であり、 $G(x) = \tanh x$ を用いた。第2章で述べたRNNはシグモイド関数のバイアスパラメータ W_m^b を有していたが、ここでは簡単のため省略している。

学習対象は、結合荷重 w をある値 w^* に固定したこのRNN、

$$q(k+1) = \hat{f}(q(k), w^*) \quad (6.2)$$

から生成される離散時系列 $q(k)$ とする。したがって、第5章の考察で仮定したように、最適点 w^* が少なくとも一つ存在することになる。

学習則は第5章の逐次型BP法(5.4)を用いる。

以下に示す数値実験では、シグモイドユニット数を $L=2$ (w の次元は $n=4$)とし、学習対象として次のようなパラメータ設定のRNNから生成される(a)~(d)の4種類の時系列を用いた(アトラクタを図6-1に示す)。すなわち、
 $w^* = (-6.0, 4.5, 1.0, W_2^{b*})$ 、

- (a) $W_2^{b*} = 2.00$: 2周期軌道、 (b) $W_2^{b*} = 2.05$: 4周期軌道、
 (c) $W_2^{b*} = 2.16$: 6周期軌道、 (d) $W_2^{b*} = 3.00$: カオス。

学習係数の値は $\eta = 0.2$ とした。この値は、(a)~(d)のすべての場合に対して第5章の式(5.20)の固有値が $|\lambda^{(1)}(k)| < 1$ の条件を満足するように選んだ。したがって、第5章の定理5-2と5-3、5-5より、学習対象が周期4以上の周期軌道の場合と概周期軌道の場合は最適点 w^* への学習の収束が、カオスの場合は学習方程式の線形近似系の収束が保証されている。

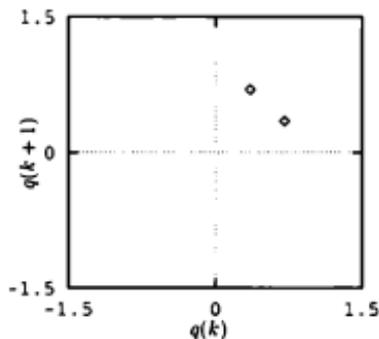
結合荷重の初期値を $w(0) = (-6.01, 4.5, 1.0, W_2^{b*})$ として上記(a)~(d)の四つの時系列をそれぞれ学習させた場合の結合荷重 w の収束の様子を (W_1^a, W_2^a) -平面および (W_1^b, W_2^b) -平面に表わし、図6-2に示した。また、第5章の定理5-1で定義した行列

$$B = Dg(w^*, M) \cdots Dg(w^*, 1)$$

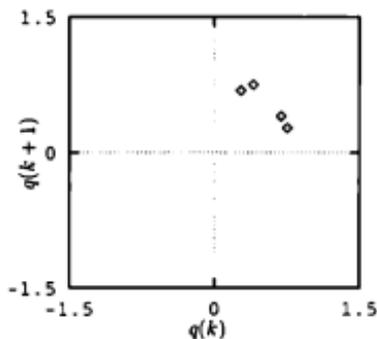
の固有値 λ および固有ベクトルを求め、その結果を特性指数 $\mu = (1/M) \log |\lambda|$ として表6-1に示した。 M は教師信号の周期である。(d)のカオスについては、不動点 w^* の法線リアプノフ指数を示している。図6-2には、周期解については、第一および第二固有値に対する固有ベクトル、カオスについては、最大法線リアプノフ指数に対するリアプノフベクトル⁶¹⁾も併せて示している。

表 6-1 不動点の特性指数と法線リアプノフ指数

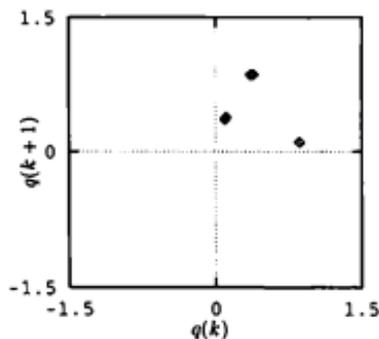
	(a)	(b)	(c)	(d)
μ_1	0.0	-6.0×10^{-6}	-1.6×10^{-9}	-4.5×10^{-3}
μ_2	0.0	-1.5×10^{-5}	-5.1×10^{-5}	-0.0085
μ_3	-0.050	-0.048	-0.066	-0.026
μ_4	-3.9	-0.84	-1.28	-1.0



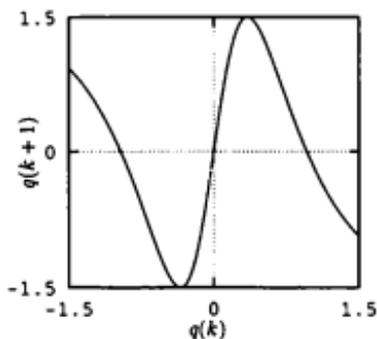
(a) $W_2^{\beta^*} = 2.00$: 2 周期解



(b) $W_2^{\beta^*} = 2.05$: 4 周期解



(c) $W_2^{\beta^*} = 2.16$: 6 周期解



(d) $W_2^{\beta^*} = 3.00$: カオス

図 6-1 学習対象のアトラクタ

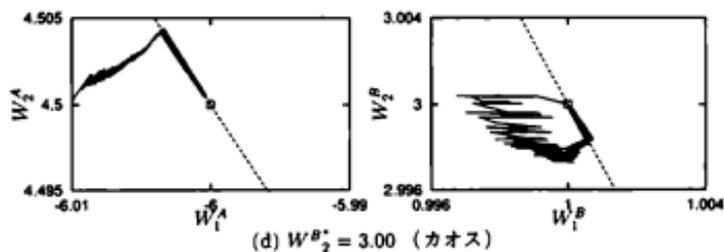
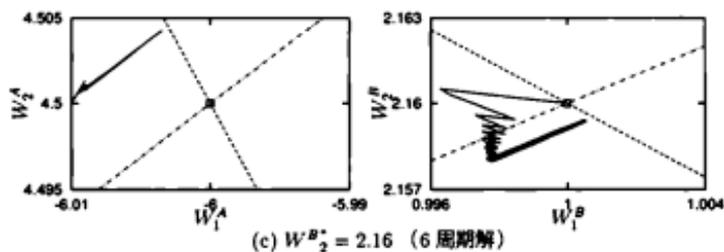
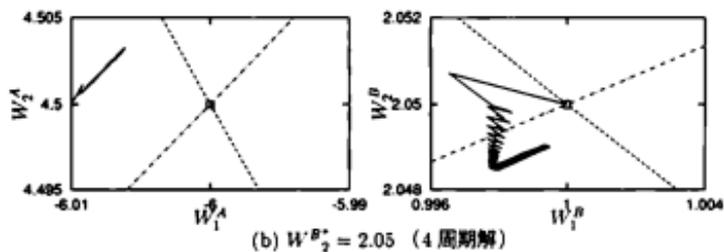
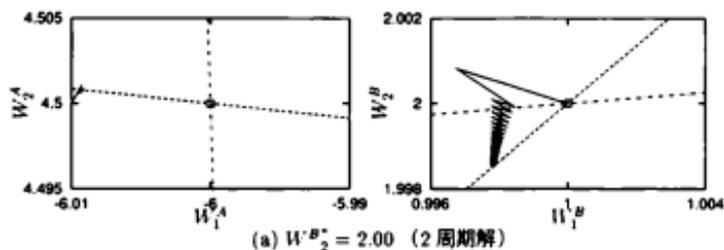


図 6-2 離散力学系学習の収束の様子

□ : 最適点 w^* 、

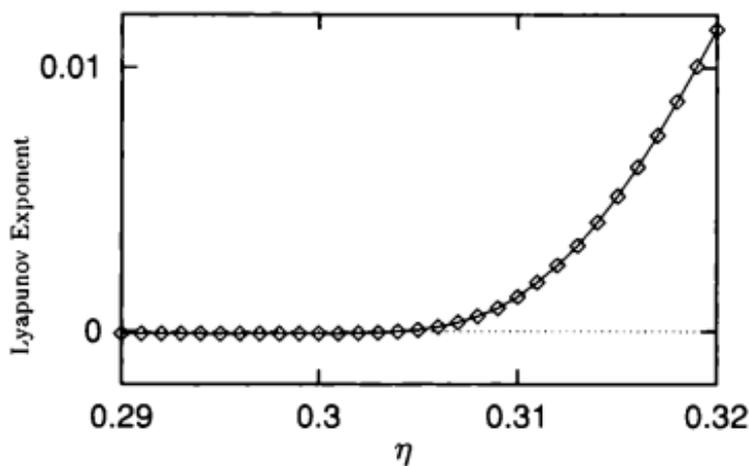
破線: 第一固有ベクトルまたは第一法線リアプノフベクトル、一点鎖線: 第二固有ベクトル

まず、(a)は2周期解で、 $2 < n$ であるから、 w^* への収束は保証されない。第5章の定理5-1の後で述べたように、 w^* を含む最適点の集合 W_{OP} が w^* の固有値1に対応する固有空間に接するように存在し、この集合上の1点に収束するものと考えられる。実際、図6-2(a)を見ると、第一固有ベクトルの方向の直線上に収束しているように見える。理論的には固有値1の固有空間は2次元であるが、数値計算上は二つの最大固有値の値に差が生じ、最適集合は1次元になっているものと思われる。

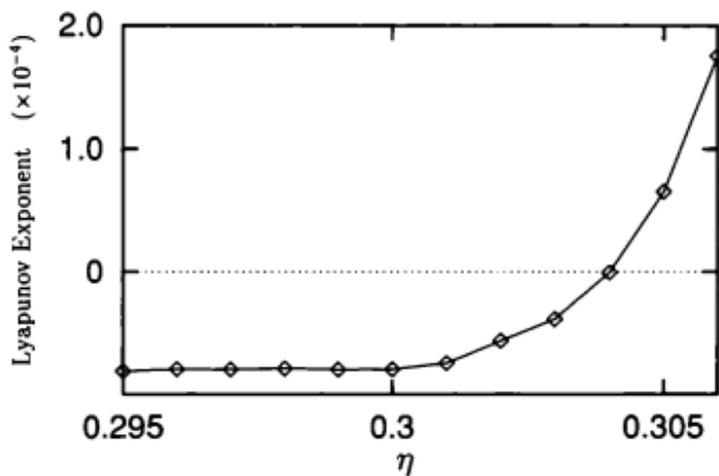
(b)と(c)はそれぞれ4周期、6周期であるから、周期 $M \geq n$ なので、定理5-1の(ii)より w^* への収束が期待される。図6-2(b)、(c)では w^* へは収束していないように見えるが、これはまだ過渡状態であって、非常に緩やかな速度で w^* に近付いているのである。表6-1に示したように、(b)、(c)とも最適点付近での学習の速度を決定する第一および第二特性指数が0に非常に近く、このために収束が非常に遅くなっているものと考えられる。0に近い値の特性指数があることは、次のように解釈できる。すなわち、理論的には最適点 w^* は孤立しているが、その近傍では誤差 $E(w, k)$ はほとんど零であり、したがってその勾配も非常に緩やかになっているのである。また、図6-2より、結合係数の変化の軌跡は第二固有ベクトルに平行して進み、第一固有ベクトルに接近していることがわかる。この後、第一固有ベクトルに沿った方向に変化し、最適点 w^* に収束するものと考えられるが、現実的な計算時間では追跡できない。(d)のカオスについては、最大法線リアプノフ指数は負であるから、線形近似系は安定であるが、学習の収束自体は保証されていない。しかし、図6-2を見ると、結合係数は最大法線リアプノフ指数に対応するリアプノフベクトルに沿って変化し、最適点 w^* に収束していることがわかる。

6.2.2 法線リアプノフ指数と最適点の安定性

図6-1(d)のカオスの学習において、学習係数 η を変化させた場合の法線リアプノフ指数の変化を図6-3に示す。この図より、最大法線リアプノフ指数は学習係数 η とともにほぼ単調に増加し、 $\eta = 0.303 \sim 0.305$ で負から正に転じることがわかる。図6-4は $\eta = 0.303$ および $\eta = 0.305$ の場合の、最適点近傍に初期値を与えた学習の収束状態を示している。前者の最大法線リアプノフ指数は -3.9×10^{-5} で負であり、後者は 6.5×10^{-5} で正である。両者の差は零をはさみ、 10^{-5} のオーダーというごくわずかなものであるが、図6-4を見ると定性的な相違が見られる。すなわち、 $\eta = 0.303$ の場合はバースト状の乱れを伴うが、最終的には学習は最適点に収束している。一方 $\eta = 0.305$ の場合はバーストが消滅せず、最適点の近傍で間欠のカオストラ



(a) $0.29 \leq \eta \leq 0.32$



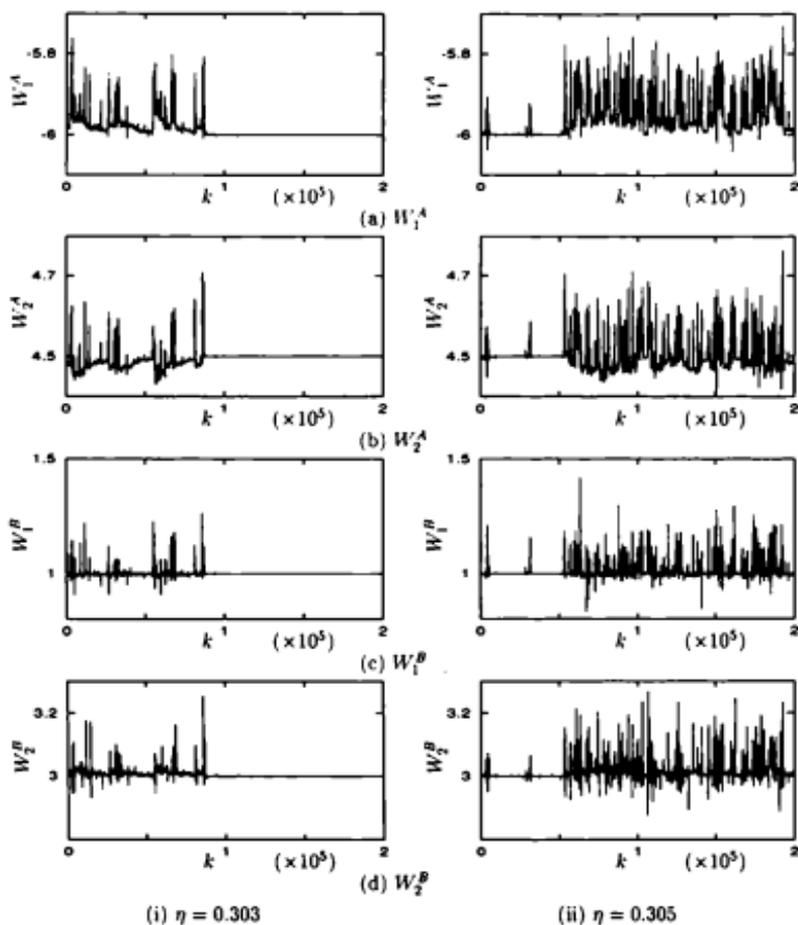


図 6-4 最適点近傍での学習の最終状態

クタになっている。

第5章で考察したように、最大法線リアプノフ指数が負であること、より一般に線形近似系の零点が(非一様)漸近安定であることは、最適点の漸近安定性を必ずしも保証しない。しかし、図6-4の例の場合は、最大法線リアプノフ指数の正負が最適点の安定性を支配していることを意味する。

また、教師信号の変化と結合荷重の変化を合わせた自律系として学習過程を捉えると、バースト状の乱れを伴う不安定化は、Ashwinら³¹⁾によって研究された“locally riddled basin”を伴うアトラクタから“chaotic saddle”への分岐であると考えられる。以下、この分岐について考察する。

まず、1次元教師信号の変化とそれに基づく結合パラメータ w の更新則を連立させると次のような $n+1$ 次元の自律的力学系となる。

$$\begin{aligned}w(k+1) &= \hat{g}(w(k), q(k)), \\q(k+1) &= \hat{f}(q(k), w^*).\end{aligned}\tag{6.3}$$

この系において、教師信号はカオスアトラクタ A_q 上で与えられるとすると、学習の最適状態は R^{n+1} の不変部分多様体 $\{w^*\} \times R$ に含まれた1次元カオス的不変集合

$$A^* = \{(w, q) \mid w = w^*, q \in A_q\}$$

で表現される。学習の収束性はこの不変集合の漸近安定性と一致する。また、法線リアプノフ指数は系(6.3)の $n+1$ 個のリアプノフ指数から教師信号のリアプノフ指数を除いたものに一致する(証明は付録に示す)。

さて、一般に微分多様体 X 上の連続写像 f で定義される離散力学系において、 A^* のように或る低次元の不変部分多様体 Y に含まれたカオス的不変集合は法線リアプノフ指数が負から正に変化するにつれて下記のような順序で変化し、安定性を喪失していくとAshwinらは主張している。

- (1) 漸近安定アトラクタ、
- (2) locally riddled basin を伴ったアトラクタ、
- (3) chaotic saddle、
- (4) normally repelling chaotic saddle。

(1)と(2)のアトラクタは以下のように定義されている。以下の定義では、集合 A はコンパクトな不変集合とする。

[定義6-1] A がリアブノフ安定であり、かつ、 A の引力圏 $B(A)$ (すなわち、極限集合が A に含まれる点の全体) が A のある開近傍を含むとき、 A は漸近安定アトラクタであるという。 □

これは、通常用いられる、アトラクタの位相的な定義である。(2) のアトラクタの定義の前に、Milnor の意味でのアトラクタの定義を述べる。

[定義6-1'] 次の二つの条件を満たす A を Milnor の意味でのアトラクタと呼ぶ⁵²⁾。

(i) $B(A)$ が正の測度を持つ。

(ii) $B(A')$ が $B(A)$ に測度零の点を除いて一致するような、 A の真部分閉集合 A' は存在しない。

ただし、多様体 X 上の測度としては、各局所座標系ではルベーク測度と等価になるものを採用する。 □

これはアトラクタの測度論的な定義である。また、 $B(A)$ が A の開近傍を含めばその測度は正であるから、漸近安定なアトラクタは Milnor の意味でのアトラクタになる。したがって、Milnor による定義は、通常のアトラクタの定義の拡張である。

[定義6-2] Milnor の意味でのアトラクタ A の引力圏 $B(A)$ に関して、 $\epsilon > 0$ が存在し、任意の $x \in B(A)$ を中心とする任意の開球 $U(x)$ について、 $U(x)$ に含まれる点のうち、それを初期値とする軌道が A の ϵ -近傍から必ず1回は出るような点の集合が正の測度を持つとき、 A は locally riddled basin を持つという。 □

locally riddled basin を持つことは、アトラクタ A のいくら近くに初期値を置いても、その軌道が A の ϵ -近傍から一旦は外に出る確率が正であることを意味し、リアブノフの意味では A は不安定である。しかし、 ϵ -近傍から外へ出た軌道が最終的には A に漸近する可能性は一般に存在する。その可能性がない場合、すなわち、その軌道が別のアトラクタに漸近するか無限遠に発散する場合、 $B(A)$ は riddled basin と呼ばれている。正確には次のように定義されている。

[定義6-2'] 任意の $x \in B(A)$ 、任意の開球 $U(x)$ について、 $U(x) \cap B(A)^c$ が、正の測度を持つとき、Milnor の意味でのアトラクタ A は riddled basin を持つという。 □

(3) と (4) の不変集合は以下のように定義されている。

[定義6-3] A はカオス的不変集合であり、かつ、 A の近傍 U があって、 $B(A) \cap U$ は

⁵²⁾ Ashwin⁵¹⁾ の定義では、条件(ii)の代わりに A が位相推移的であること、すなわち、 A が稠密な軌道を含むという条件を課している。二つの条件はともにアトラクタがある意味で極小であることを保証するものであるが、両者は微妙に異なり、一般には互いに独立である。ここで述べている、不変集合に関する種々の定義はすべて Ashwin⁵¹⁾ によるが、Milnor の意味でのアトラクタについては Milnor 自身による定義を採用する。

A より大きいとその測度が零のとき、 A を chaotic saddle と呼ぶ。 □

[定義 6-4] A が成る低次元の不変部分空間 Y に含まれた chaotic saddle であり、さらに $B(A)$ も Y に含まれてしまうとき、 A を normally repelling chaotic saddle と呼ぶ。 □

したがって、normally repelling chaotic saddle A は Y に横断的な方向には完全に不安定である。

さて、法線リアブノフ指数が負から正に変わるときに、locally riddled basin を持つアトラクタが chaotic saddle に変化すると考えられている。これは、次のように説明される。

まず、一般に不変集合 A のリアブノフ指数は A の近傍の軌道の平均的な発散率を表わす。ここで、平均は A の「自然な」不変測度に関して定義されている。すなわち、 A の不安定多様体上で定義されたルベーク測度に関して絶対連続な測度での平均である³¹⁾。自然な測度が物理的・数値計算的に観測される唯一の測度であるが、カオスアトラクタの中には一般に無数の不安定周期軌道が含まれており、それぞれに対応した不変測度が存在する。したがって、法線リアブノフ指数が負であっても、不安定周期軌道に対応する不変測度で平均した軌道の発散率は正になっている可能性がある。また、カオス軌道は不変集合 A において稠密であるから、不安定周期軌道の任意の近傍を通過する。このように、正の発散率を持つ不安定周期軌道に接近するとき、 A の近傍の軌道は一時的に A から遠ざかる。これが locally riddled basin の生じる理由である。また、法線リアブノフ指数が正になると自然な不変測度で平均した発散率が正であるから、 A の近傍の軌道は確率 1 で A から遠ざかっていく。すなわち、不変集合 A は chaotic saddle に変化するのである。

そこで、学習を表わす系 (6.3) に戻って考えると、部分空間 $\{w^*\} \times R$ に含まれない、 A^* の近傍の軌道は平均的には A^* に漸近するが、不安定周期軌道の近傍を通過するときに A^* から離れていく可能性がある。実際、 $\eta = 0.303$ のとき、学習対象系 (6.2) は不安定不動点 $\hat{q} \approx 0.7101512686$ を持ち、その最大法線リアブノフ指数は 0.452 で正である。したがって、教師信号が不動点の近傍を訪れるとき、一時的に学習が不安定になり、バースト状の乱れが現れるのである³²⁾。不動点 \hat{q} の逆像は A_q の中で稠密と考えられるから、locally riddled basin の条件を満たしているものと思われる。しかし、 A^* から離れていった軌道は他のアトラクタに漸近することなく、最終的には A^* に収束することが初期値を変えた実験により確認された。よって、 A^* の引力圏は riddled basin にはなっていないものと考えられる。

³¹⁾ Ashwin らはこのバーストを “bubbling” と呼んでいる³¹⁾。

riddled basin になるか、単に locally riddled basin であるかは、系の大域的構造によって定まる。たとえば、不変部分空間 X に独立な成分（学習の場合は w ）に関して系が線形ならば、 A^* 中の不安定不動点は無限遠に連絡する不安定多様体を持ちうる。しかし、リアプノフ指数が負であれば系の線形性より $B(A^*)^C$ は測度零になり、riddled basin は生じない。このように、riddled basin の発生は非線形性が本質的に関与するものであり、一般的な解析は困難である。

一方、 $\eta = 0.305$ になると、法線リアプノフ指数は正になり、 A^* は安定性を喪失して chaotic saddle となる。すなわち、 A^* に漸近するルートが存在したとしても、それは鞍形点の安定多様体のように測度零である。このとき A^* の近傍の軌道の収束先は A^* を含む間欠的カオスアトラクタになっている。 A^* が安定性を失った後のアトラクタが A^* を含む間欠性カオスになることは、分岐前の η の値においてもアトラクタは A^* のみであり、共存するアトラクタがないことを示唆している。したがって、このことから、分岐前の A^* は riddled basin を持たないことがわかる。

6.2.3 最適点が不安定化したときの学習の最終状態

本節では、学習係数 η を大きくして不動点 w^* が不安定になった場合に、結合係数がどのような状態に収束するかを調べる。学習対象は (d) のカオス時系列とする。

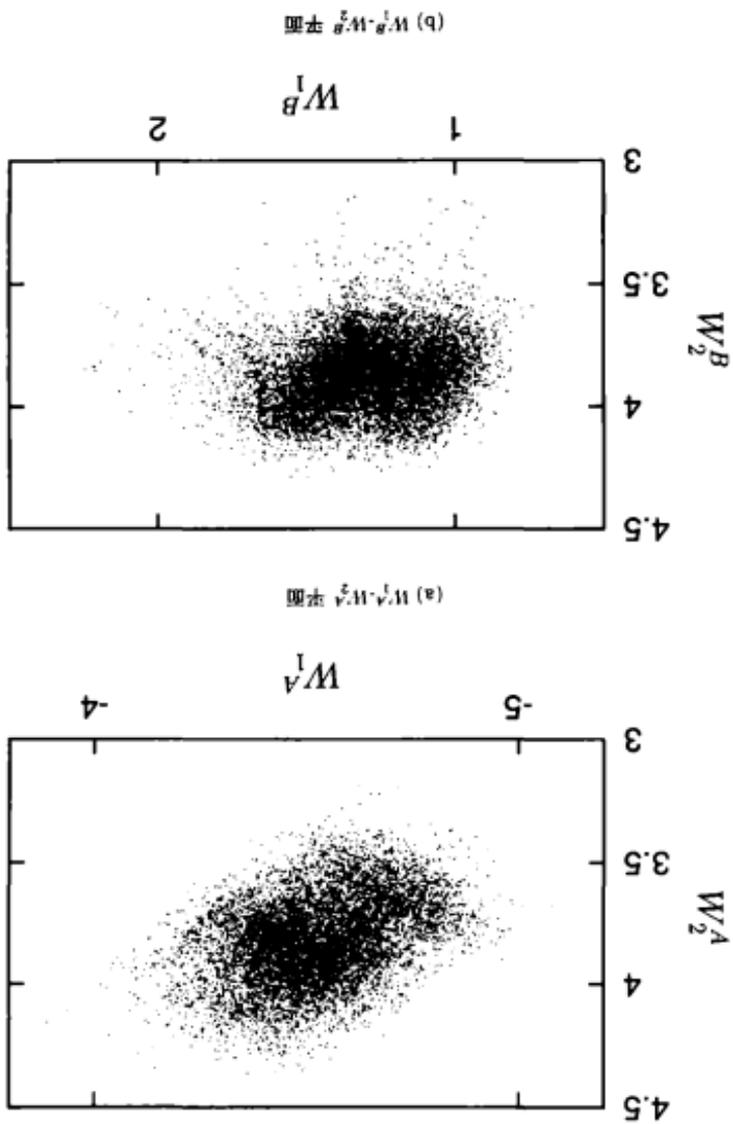
力学系 (5.4) はカオス的な外力によって駆動される非自律系であるから、不動点が不安定になった場合には、定常状態はカオスであると予想される。

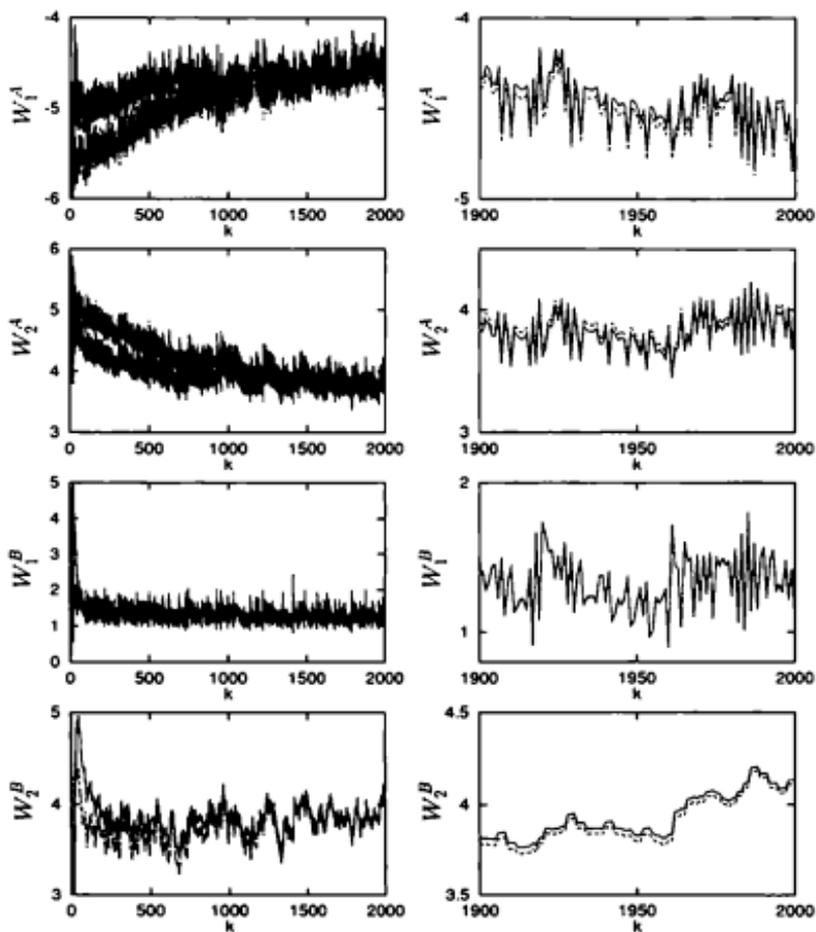
図 6-5 は、初期値を $w(0) = (-6.01, 4.5, 1.0, 3.0)$ 、学習係数を $\eta = 0.5$ とし、前節 (d) のカオス時系列を学習させた場合の結合係数 w の最終状態を示している。不動点 w^* の最大法線リアプノフ指数は $\mu = 0.46$ であるからこの不動点は不安定になり、定常状態は予想通りカオス状になっている。

ところが、この定常状態の最大法線リアプノフ指数（式 (5.26) で w^* を $w(k)$ とする）を測定すると、 $\mu = -0.0016$ という負の値になっていた。これは、この状態が「カオスではあるが軌道安定である」ことを示唆している。

このことを確認するため、異なる二つの初期状態から始めた学習による W_m^A および W_m^B の時間変化を計算し、図 6-6 に示した。図 6-6(b) は図 6-6(a) の最終状態の拡大であり、二つの波形は同じ状態に漸近していることが見てとれる。以上の結果は、このカオス状態はあくまでカオス的な外力によって駆動されたものであり、自律的に発生したものではないことを意味している。

図6-5 結合荷重のカーブの挙動





(a) $0 \leq t \leq 2000$

(b) $1900 \leq t \leq 2000$ (拡大図)

図 6-6 異なる初期値による二つの学習の収束過程

6.3 勾配降下法による力学系の学習に現れる riddled basin

6.3.1 riddled basin を持つ学習系の構成と存在証明

上述の RNN による 1 次元離散力学系の学習実験では、法線リアプノフ指数が負であっても教師信号であるカオスアトラクタに含まれる不安定周期軌道を起点とする不安定なルートが存在するため、 $n+1$ 次元自律系に拡張した空間での最適集合は locally riddled basin を持つことを示した。しかし、riddled basin は観測されていない。つまり、最適集合から一旦離れる軌道も最終的にはそこに漸近し、したがって学習は収束する。第 5 章で述べたように、線形近似によるカオス学習の収束の判定は困難ではあるが、この場合は、法線リアプノフ指数による安定性の評価が妥当であったことになる。

RNN による学習実験では riddled basin は観測されなかったが、locally riddled basin が生じていることは、力学系の学習において riddled basin が生じる可能性を示唆している。locally riddled basin が riddled basin になるか否かはアトラクタ近傍の大域的性質により定まるからである。そこで、本節では、逐次型の勾配降下法による力学系の学習においてその最適点の引力圏が riddled basin になるような例の構成を試みる。

まず、学習対象としてテント写像、

$$q(k+1) = T(q(k)) = \begin{cases} 2q(k) & (0 \leq q \leq \frac{1}{2}) \\ 2-2q(k) & (\frac{1}{2} < q \leq 1) \end{cases} \quad (6.4)$$

を考え、これを 1 次元のパラメータ w を含む系、

$$x(k+1) = \hat{f}(x(k), w) = (g(w) + 2) \times \begin{cases} x(k) & (0 \leq x \leq \frac{1}{2}) \\ 1 - x(k) & (\frac{1}{2} < x \leq 1) \end{cases} \quad (6.5)$$

$$g(w) = \begin{cases} -\frac{1}{3}(w+4) & (w < -1) \\ w & (|w| < 1) \\ \frac{1}{3}(-w+4) & (w > 1) \end{cases} \quad (6.6)$$

で学習する。学習は誤差評価関数、

$$E(w, q(k)) = \frac{1}{2} [\hat{f}(q(k), w) - T(q(k))]^2 = \frac{1}{8} g(w)^2 T(q(k))^2 \quad (6.7)$$

を最小化する勾配降下法、

$$w(k+1) = w(k) - \frac{1}{4} \eta \cdot g(w(k)) \cdot g'(w(k)) \cdot T(q(k))^2 \quad (6.8)$$

により実行する。誤差評価関数 $E(w, q)$ は式 (6.7) のように w に依存する部分と q に依存する部分の積の形で表わされる。このうち、 w に依存する部分、

$$8E(w, q(k))/T(q(k))^2 = g(w)^2$$

のグラフの形状を図 6.7 に示す。学習の最適点は式 (6.5) より明らかに $g(w)$ の零点であるから、0 と ± 4 である。これらは、式 (6.8) を結合荷重 w を変数とする離散力学系と考えた場合、その不動点である。ここでは、 $w^* = 0$ 、 $w^* = +4$ 、 $w^* = -4$ と表わし、 w^* の引力圏を考える。

まず、不動点 $\bar{w} = w^*$ または w^* における、式 (6.8) の線形化方程式は、

$$w(k+1) = \left[1 - \frac{1}{4}\eta \cdot g'(\bar{w})^2 \cdot T(q(k))^2\right]w(k) \quad (6.9)$$

である。また、テント写像はルベーク測度を保ち、かつ $q = 1/2$ に関して対称であることから、法線リアブノフ指数は次のように解析的に計算できる。

$$\lambda_1 = 2 \int_0^1 \log|1 - ax^2| dx$$

である。ただし、 $a = \eta g'(w^*)^2$ とおいた。よって、

$$\lambda_1 = \begin{cases} (1 + \frac{2}{\sqrt{a}})\log(1 + \frac{\sqrt{a}}{2}) + (1 - \frac{2}{\sqrt{a}})\log(1 - \frac{\sqrt{a}}{2}) - 2 & (a \leq 4) \\ (1 + \frac{2}{\sqrt{a}})\log(1 + \frac{\sqrt{a}}{2}) + (1 - \frac{2}{\sqrt{a}})\log(\frac{\sqrt{a}}{2} - 1) - 2 & (a > 4) \end{cases} \quad (6.10)$$

である。以上の準備より、2次元力学系

$$\begin{pmatrix} w(k+1) \\ q(k+1) \end{pmatrix} = F \begin{pmatrix} w(k) \\ q(k) \end{pmatrix} = \begin{pmatrix} w(k) - \frac{1}{4}\eta \cdot g(w(k)) \cdot g'(w(k)) \cdot T(q(k))^2 \\ T(q(k)) \end{pmatrix} \quad (6.11)$$

のカオス的不変集合

$$A = \{(w, q) \mid w = w^*, 0 \leq q \leq 1\}$$

が riddled basin を持つことが示せる。

[定理 6-1] 系 (6.11) の不変集合 A が riddled basin を持つアトラクタとなるような学習係数 η の値が存在し、かつ、そのような値の集合は η の空間 R^+ の中で開集合を含んでいる。

(証明) まず、線形化方程式 (6.9) の係数を

$$\mu(\eta, q, w) = 1 - \frac{1}{4}\eta g'(w)^2 T(q)^2 \quad (6.12)$$

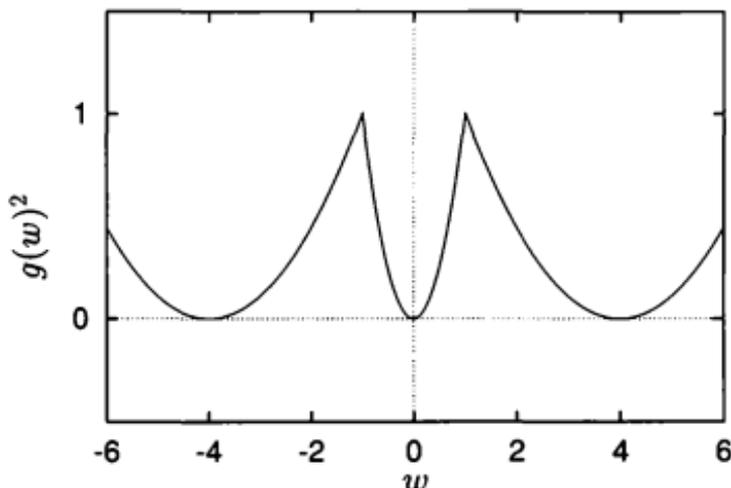


図 6-7 誤差評価関数 : $g(w)^2 = 8E(w, q)/T(q)^2$

と表わす。テント写像の不動点 $q = 2/3$ において、

$$\mu(\eta, 2/3, w^*) = 1 - \frac{\eta}{9}$$

なので、 $\eta > 18$ ならば、 $\mu(\eta, 2/3, w^*) < -1$ となる。

そこで、以下、 $\eta = 22$ の場合を考えこの値およびその近傍で A が riddled basin を持つことを示す。

まず、式(6.10)と $4.6 < \sqrt{22} < 4.7$ の関係より $\eta = 22$ のとき、

$$\begin{aligned} \lambda_1 &= \left(\frac{2}{\sqrt{22}} + 1\right) \log\left(\frac{\sqrt{22}}{2} + 1\right) + \left(1 - \frac{2}{\sqrt{22}}\right) \log\left(\frac{\sqrt{22}}{2} - 1\right) - 2 \\ &< \left(\frac{2}{4.6} + 1\right) \log\left(\frac{4.7}{2} + 1\right) + \left(1 - \frac{2}{4.7}\right) \log\left(\frac{4.7}{2} - 1\right) - 2 \\ &< 1.44 \times 1.25 + 0.58 \times 0.31 - 2 = -0.0202 < 0 \end{aligned}$$

となり、文献(48)の補題により、 $B(A)$ の測度は正であることがわかる¹。さらに、 A で稠密となるようなテント写像の軌道の初期値集合は測度1であるから、 A は定義6.1の条件(ii)も満たす。したがって、不変集合 A はMilnorの意味でのアトラクタである。

また、

$$1 - \frac{\eta}{36} \leq \mu(\eta, q, \pm 4) \leq 1 \quad (\eta \leq 36, 0 \leq q \leq 1).$$

であるから、 $7/18 \leq \mu(22, q, \pm 4) \leq 1$ である。さらに、 $|w| > 1$ の領域では系(6.8)は w に関して線形であるから、定理5.5より $|w| > 1$ で w^+ にも w^- にも軌道が漸近しない初期値集合は測度零であることがわかる。

以上のことから、 $B(A)$ がriddled basinであることを証明するには、軌道が $|w| > 1$ に入るような初期値集合が $|w| \leq 1$ の領域において開かつ稠密な集合を含むことを示せばよい。なぜならば、 $B(A)$ の点のうち、 $|w| > 1$ の領域のものについては、上述のことからその任意の近傍に含まれる $B(A)^c$ の点の集合は測度正になることがわかる。また、 $|w| \leq 1$ の領域に含まれる点については、その任意の近傍に含まれる点のうち、軌道が $|w| > 1$ に入るような初期値集合を V とすると、 V は内点を持つので測度正である。また、 V は q 軸と交わらないと仮定できるから、ある有限の k について $F^k(V)$ は $|w| > 1$ の領域に入る。テント写像は正方向には測度を2倍にし、また f は w に関して区分別形であるから、 $F^k(V)$ は測度正であり、かつ F^k は測度零の集合を測度正の集合に写さない。さらに $F^k(V) \cap B(A)^c$ は測度正であるから、結局 $V \cap F^{-k}(B(A)^c) = V \cap B(A)^c$ も測度正となる。

さて、

$$\lambda = \mu(\eta, 2/3, w^+) = 1 - \eta(1 - 2/3)^2 < -1$$

と仮定する。このとき、 $\lambda' = \lambda + \epsilon < -1$ かつ

$$0 < \delta \leq \frac{1}{3} - \sqrt{\frac{1}{9} - \frac{\epsilon}{\eta}} < \frac{1}{6}, \quad \frac{1}{9} - \frac{\epsilon}{\eta} > 0 \quad (6.13)$$

となる $\epsilon > 0$ と $\delta > 0$ が存在すれば、

$$\mu(\eta, q, w^+) \leq \lambda' \quad (2/3 - \delta \leq q \leq 2/3 + \delta)$$

¹ $\log(4.7/2+1) < 1.25$ と $\log(4.7/2-1) < 0.31$ であることは、 $\exp(1.25)$ と $\exp(0.31)$ のテラー展開の有限項をそれぞれ計算することにより、厳密に示せる。

² 文献(48)の証明には不備があるように思われるが、今の例のように系が区分別形の場合には補題は成立している。証明は付録に記す。

となる。また、

$$N > \frac{-\log |u'|}{\log |\lambda'|} \quad (6.14)$$

かつ、

$$|q(0) - \frac{2}{3}| \cdot 2^{N+1} \leq \delta \quad (6.15)$$

であれば、

$$|u'| |\lambda'|^N \geq 1$$

かつ、

$$|q(k) - \frac{2}{3}| \leq \delta \quad (0 \leq \forall k \leq N+1)$$

となる。よって、領域

$$D_0 = \{ (w, q) \mid |w| > (\frac{2|q - 2/3|}{\delta})^{\log_2 |\lambda'|} \}$$

に初期値を与えると、軌道は $|w| > 1$ の領域に入る。

そこで、

$$D = \bigcup_{k=0}^{\infty} F^{-k}(D_0)$$

とすると、 D は軌道が $|w| > 1$ の領域に入る初期値集合に含まれる開集合である。

また、

$$C_k = \{ (w, q) \mid w \neq 0, q = T^{-k}(2/3) \}$$

とすると、 $C_k \subset F^{-k}(D_0)$ であり、かつ、 $\bigcup_{k=0}^{\infty} C_k$ は $|w| \leq 1$ の領域で稠密であるから、 D は開かつ稠密な集合になる。

そこで、 $\eta = 22$ の場合を考えると、

$$\lambda = \mu(22, \frac{2}{3}, w^*) = -\frac{13}{9} < -1$$

であり、 $\epsilon = 1/3$ とすると、

$$\lambda' = \lambda + \epsilon = -\frac{10}{9} < -1$$

で、

$$\delta \leq \frac{1}{3} - \sqrt{\frac{1}{9} - \frac{\epsilon}{\eta}} = \frac{1}{3} (1 - \sqrt{\frac{19}{22}}) < \frac{1}{6}$$

であるから、 $\delta < 1/6$ の条件も満たしている。

以上より、系 (6.11) の不変集合 A は、 $\eta = 22$ のとき、riddled basin を持つアトラクタとなることが示せた。また、 ϵ を

$$0 < \epsilon < \min(\eta/12, -\lambda - 1)$$

のように選んでおけば、条件(6.13)を満たす δ は必ず存在する。よって、 A が riddled basin を持つパラメータ η の値は22で孤立しておらず、これを内点として含んでいることがわかる。よって定理は証明された。□

図6-8に、 $\eta = 22.0$ の場合のアトラクタ A の引力圏の数値計算例を示す。ただし、数値計算のためテント写像の係数は $2.0 - 10^{-10}$ とした。学習系(6.11)は $w = 0$ に関して対称であるから、図6-8では $w \geq 0$ の領域のみ示している。白地が A の引力圏 $B(A)$ を表わしている。また、黒く塗りつぶした領域が $\{w^+\} \times [0, 1]$ の、灰色の領域が $\{w^-\} \times [0, 1]$ の引力圏をそれぞれ表わしており、その和集合が $B(A)^c$ である。 $B(A)^c$ は A の近傍では極めて幅が狭いため、稠密であることは確認しづらいが、引力圏は理論通り riddled basin になっていることがわかる。

6.3.2 学習係数をカオス的に動振した勾配降下法

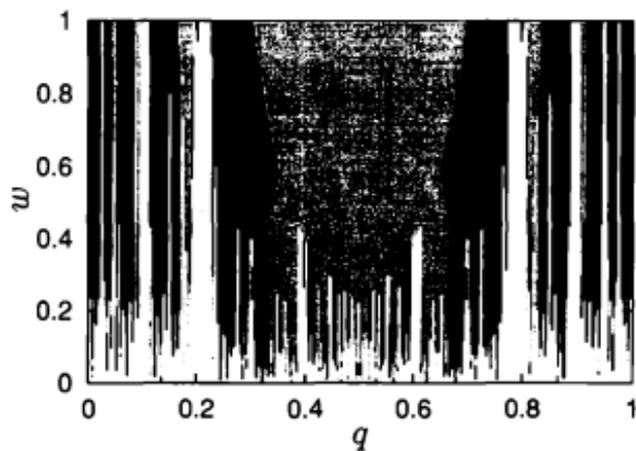
式(6.8)の勾配降下法を

$$w(k+1) = w(k) - \left(\frac{1}{4}\eta \cdot T(q(k))^2\right)g(w(k)) \cdot g'(w(k))$$

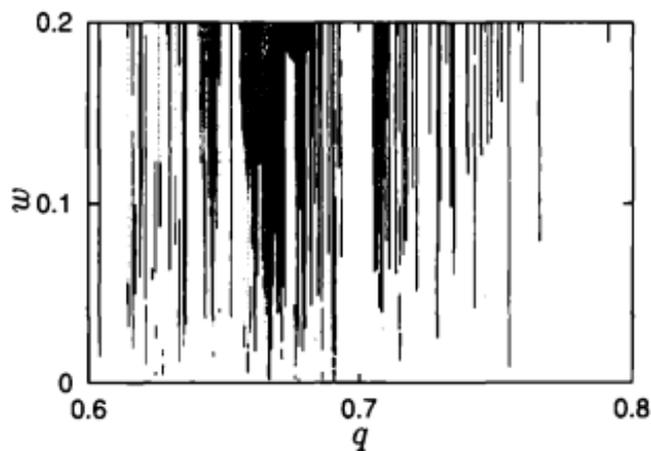
と表わすと、これは評価関数 $(1/2)g(w)^2$ を最小化する勾配降下法で、学習係数を $(1/4)\eta \cdot T(q(k))^2$ のようにカオス的に動振したものに一致することがわかる。このことから、riddled basin を勾配降下法におけるローカルミニマムの回避に利用することが考えられる。すなわち、カオス的なノイズを加えることにより、ローカルミニマムの引力圏の測度を制御し、学習がそこに捕捉される確率を小さくする方法である。カオスは単なるランダムノイズではないから制御方法がある程度設計可能で、かつ、ノイズは学習係数に加えるのであるから勾配降下方向には影響を与えないという利点がある。また、カオス的な信号、そのアトラクタに含まれる不安定周期軌道に起因する短期時間相関のため、学習系にノイズとして加えた場合、ローカルミニマムからの脱出確率がランダムノイズを加えた場合よりも高いという報告⁶³⁾や、周期軌道の学習においてNNが生成するカオスがローカルミニマムからの脱出に有効であるという報告⁶⁴⁾ともこの方法は密接に関係するものと思われる。

6.4 結言

第5章に引き続き、逐次型のBP法のような二乗誤差最小化の勾配降下法による力学系の学習を学習系のシステムパラメータを変数とする力学系の状態変化と



(a) $0.0 \leq q \leq 1.0, 0.0 \leq w \leq 1.0$



(b) $0.6 \leq q \leq 0.7, 0.0 \leq w \leq 1.0$

図 6-8 不変集合 A の引力圏

して考察した。本章では、簡単なRNNによる1次元離散力学系の学習実験を行ない、結合荷重の最適点への収束の様子を観測した。

数値実験の結果、周期解を学習する場合は理論どおり最適点（あるいは最適集合）に収束するが、その収束の速度は非常に遅いことがわかった。また、カオスを学習する場合には、今回の実験で対象とした系では法線リアプノフ指数が負であれば最適点は漸近安定になっていた。しかし、法線リアプノフ指数が負であっても、教師信号であるカオス軌道が、その閉包に含まれる不安定周期軌道の近傍を通過するとき、最適点が一時的に不安定になることがあり、この場合には、最適点への結合荷重の収束過程にバースト状の乱れを伴うこと、すなわち、最適点の引力圏がlocally riddled basinになることを明らかにした。

また、カオスの学習において最適点が不安定になったときの結合荷重の変化の最終状態を観測し、カオス的な外力に駆動された軌道安定なカオス状態になっていることを示した。

さらに、逐次型勾配降下法による離散力学系の簡単なモデルを構成し、その学習の最適点の引力圏がriddled basinになること、また、そのような引力圏の生じるような学習係数 η の集合が開集合を含むことを証明し、数値実験によりこれを確認した。

第7章 結論

7.1 本研究の成果

本論文では、階層型NNによる力学系の学習について、学習過程に生じる現象の分析と学習過程を結合荷重空間上の力学系とみなした場合の解析を行なった。以下、本研究で得られた成果を要約する。

(1) RNNにおける学習則として、離散力学系を対象とした関数誤差法と軌道誤差法と、連続力学系を対象とした速度誤差法と波形誤差法を導き、本研究では扱わない他の学習法も含めて、各種学習則の特徴を整理した。

(2) 確率的降下法の理論をRNNの学習則の解析に適用し、本研究で扱うRNNの学習則がすべて確率的降下法として定式化されることを示した。また、離散力学系の関数誤差法と連続力学系の速度誤差法がBP法に平均的な意味で等価であることを証明し、学習係数と誤差評価時間の大きさから学習速度を評価する式を導出した。

また、甘利による学習収束の定理を連続力学系の適応制御型学習に拡張し、空間のコンパクト性を仮定して議論を厳密化するとともに、このかたちの学習が最適集合に確率1で収束することを証明した。

(3) 1次元離散力学系の学習実験を行ない、周期軌道および不安定性の小さいカオスの学習においては、BP法と関数誤差法が平均的には等価であることを示し、確率的降下法の理論に基づく学習速度の評価が妥当であることを確認した。また、前記の学習実験においてカオス学習の加速現象を発見し、その原因が逐次型の勾配降下法の持つ確率的性質とカオスの軌道不安定性がシグモイド関数のバイアス項に揺らぎを与えることにより生じていることを明らかにし、対称性を持つ系に特有の性質であることを示した。

(4) 離散力学系の学習において、教師信号の提示方法が学習に及ぼす影響を数値実験により考察し、1次元系の学習においてはデータ数が20程度でデータ数無制限の場合と同程度の学習が可能であることを、教師信号の分布密度が学習に大きく影響し、これに比べて時間相関の及ぼす影響は小さいことを示した。

(5) 連続力学系の学習においても、学習初期の速度は確率的降下法の理論に基づく評価式に従うことを確認した。

(6) 逐次型のBP法を含む二乗誤差最小化の勾配降下法による力学系の学習をシステムパラメータを変数とする非自励的力学系の状態変化としてとらえ、最適点近傍での学習の収束条件を不動点の安定性として解析的に考察した。

まず、離散力学系の逐次型BP法による学習の最適点への収束性に関して、以下のことを証明した。周期軌道の学習は、学習対象の次元と周期の積がパラメータ空間の次元より大きい場合には局所収束する。非周期的な概周期軌道の学習は一般に局所収束する。カオス軌道の学習は、線形近似系の零点の一樣漸近安定性が一般には成立しないため、一般的な収束条件は明らかでないが、線形近似系の零点の漸近安定性は確率1で保証される。したがって、結合荷重に関して線形なRNNでカオスを学習する場合には、一般的な仮定のもとに学習は最適点に収束する。この結果は関数誤差法・軌道誤差法による学習にもほぼそのまま拡張できる。

また、連続力学系の最適制御型学習則による学習の収束性に関しては、以下のことを証明した。速度誤差法による学習では、学習対象のアトラクタが、周期軌道、概周期軌道の場合は一般に、平衡点の場合はパラメータの個数が学習対象の系の次元より小さければ局所収束が保証される。また、カオスの学習については、離散系の学習と同様、線形近似解析が困難であり、線形なRNNによる学習の収束性のみ保証される。この結果は波形誤差法による学習の場合も、ほぼそのまま成立する。

(7) 結合荷重が4個の簡単なRNNによる1次元離散力学系の学習実験を行ない、周期解を学習する場合は非常に緩慢な速度で最適集合に収束すること、また、カオスを学習する場合には、法線リアプノフ指数が負であれば最適点は漸近安定になっていることを示した。しかし、法線リアプノフ指数が負であり最適点が漸近安定であっても、その引力圏がlocally riddled basinになりうることを明らかにした。また、カオスの学習において最適点が不安定になったときの結合荷重の変化の最終状態を観測し、カオス的な外力に駆動された軌道安定なカオス状態になっていることを示した。

さらに、逐次型勾配降下法による離散力学系の簡単なモデルを構成し、その学習の最適点の引力圏がriddled basinになること、また、そのような引力圏の生じるような学習係数 η の集合が開集合を含むことを証明し、数値実験によりこれを確認した。

7.2 今後の課題

(1) カオスの学習の収束条件の厳密化

本研究では、力学系の学習の最適状態への収束条件を、教師信号が周期的、概周期的、カオス的な各場合についてそれぞれ考察した。このうち、カオスの学習に関しては、線形近似系が確率1で収束するための一つの条件を得たが、一般的な解析はまだ十分でない。解析の困難さは、学習系がカオス時系列によって駆動された非自律的力学系であることに起因する。カオス学習の収束条件をより精密にすることは、カオスの外力を持った系の安定性解析という、非線形力学の新しい問題提起という点からも有意義であると考えられる。

(2) カオスの非極小性の応用

カオス学習の収束性の解析において線形近似が成立しないことや、最適状態が locally riddled basin になることなどは、カオスが不安定周期軌道を無数に含むこと、すなわちカオスアトラクタの非極小性に基づくものである。非極小性は、初期値に対する鋭敏な依存性や軌道不安定性、混合性、フラクタル性とともにかオスの持つ重要な特徴の一つであるが、従来あまり注目されていなかった。しかし、近年、カオスアトラクタに含まれた不安定周期軌道を安定化制御する手法が提案され、微小な制御操作により多様な信号を出力するシステムを構築可能にする技術として期待されている。第6章で触れた riddled basin の学習系への応用も含め、カオスの非極小性の工学的利用が今後の重要なテーマになるであろう。

上記2点を含めたカオスの基礎理論、工学的応用面での研究が発展するとともに、本研究で得られた知見がNNによる時系列予測・システム同定技術の向上に活用されることを期待する。

謝 辞

本研究は京都大学工学部上田 皖亮教授の御指導のもとに遂行したものである。終始適切な御教示と御鞭撻を賜りましたこととともに、「カオス動力学」という研究分野を創始・開拓いただきましたことに対し、深甚なる謝意を表します。

また、本研究を遂行するに当たり、多くの有益な御助言と御援助を賜りました京都大学工学部高瀬冬人助手、名古屋大学工学部松村年郎教授に深く感謝いたします。

関西大学工学部引原隆士助教授には、筆者の京都大学工学部四回生当時から、多くの御指導・御援助を賜りました。ここに衷心より謝意を捧げます。

京都大学工学部奥村浩士教授、倉光正己講師、喜多一助手、京都府立山城高等学校太田博文教諭には有益な御助言を賜わるとともに、終始暖かい御激励を賜りました。ここに深く感謝の意を表します。

また、長尾茂樹、池上史彦、甲田哲也、久保博、大西隆史、藤井丈仁、Peachawat Peachavanish の各氏には、京都大学大学院および京都大学工学部における特別研究として多大なる御協力を頂きました。ここに厚く御礼申し上げます。

最後に多くの御援助を賜りました上田研究室の諸氏、有益な御助言・御討論を賜りました大阪大学工学部潮俊光助教授ならびに関西非線形問題懇親会（お座敷）の各位に厚く御礼申し上げます。

付録 A リカレントニューラルネットワークの学習則の導出

第2章で述べたRNNによる力学系の学習則の導出過程を述べる。連続力学系の速度誤差法は村上・佐藤²³⁾によって導出されたものであり、波形誤差法および離散力学系の二つの学習則はこれを拡張したものである。以下、用語・記号は村上・佐藤²³⁾に準ずる。

第2章ではすべてのダイナミックユニット DU に教師信号が与えられる場合の学習則を述べたが、ここでは、 DU の一部にのみ教師信号が与えられるような、より一般的な設定の場合の学習則を導出する。

まず、 DU のうち、教師信号の与えられるユニットをヴィジブル・ダイナミックユニット (VD) と呼び、与えられないユニットを隠れダイナミックユニット (HD) と呼ぶことにする。すなわち、一般に、 $DU = VD \cup HD$ かつ、 $VD \cap HD = \emptyset$ である。第2章で扱ったのは $DU = VD$ の場合である。

A.1 離散力学系の学習則

上記の一般的な設定の場合、関数誤差法と軌道誤差法の誤差評価関数は、それぞれ次のようになる。

$$E_1(\mathbf{w}) = \frac{1}{2} \sum_{k=0}^{N_B-1} \sum_{i \in VD} [\hat{f}_i(\mathbf{x}(k), \mathbf{w}) - f_i(\mathbf{q}(k))]^2, \quad (\text{A.1})$$

$$E_2(\mathbf{w}) = \frac{1}{2} \sum_{k=0}^{N_B-1} \sum_{i \in VD} [x_i(k+1) - q_i(k+1)]^2. \quad (\text{A.2})$$

ただし、関数誤差法において、 $\mathbf{x}(k)$ は

$$x_i(k) = q_i(k) \quad (i \in VD, k = 0, \dots, N_B - 1)$$

のように VD の成分を教師信号に拘束したRNNの出力であり、一方、軌道誤差法において、 $\mathbf{x}(k)$ は VD の初期値のみを

$$x_i(0) = q_i(0) \quad (i \in VD)$$

のように教師信号に一致させた場合のRNNの出力である。両方法とも、教師信号は VD にのみ与えられるので、 HD のユニットについては、その初期値 $x_i(0)$ ($i \in HD$)

も更新する必要がある。そこで、誤差評価関数 E_1 、 E_2 を減少させるべく、次の勾配降下法に基づき、パラメータ w と $x_i(0)$ を更新することにより学習を行なう。

$$w \leftarrow w - \eta \nabla_w E_l \quad (l = 1, 2). \quad (A.3)$$

$$x_i(0) \leftarrow x_i(0) - \eta \frac{\partial E_l}{\partial x_i(0)} \quad (l = 1, 2, i \in HD). \quad (A.4)$$

A.1.1 関数誤差法

まず、

$$J_i(k) = \hat{f}_i(x(k), w) - f_i(q(k)) \quad (i \in VD) \quad (A.5)$$

とすると、誤差評価関数 E_1 は次のように表現できる。

$$E_1(w) = \sum_{k=0}^{N_g-1} \left\{ \frac{1}{2} \sum_{i \in VD} J_i(k)^2 - \sum_{i \in HD} p_i^X(k) [x_i(k+1) - \sum_{m=1}^M W_{im}^A G(y_m(k))] - \sum_{m=1}^M p_m^Z(k) [G(y_m(k)) - G(\sum_{j=1}^n W_{mj}^B x_j(k) + W_m^D)] \right\}. \quad (A.6)$$

そこで、パラメータ w の修正に対する E_1 の変分を計算すると、

$$\begin{aligned} \delta E_1(w) &= \sum_{k=0}^{N_g-1} \sum_{i \in VD} J_i(k) \sum_{m=1}^M \delta W_{im}^A G(y_m(k)) + \sum_{k=0}^{N_g-1} \sum_{i \in HD} p_i^X(k) \sum_{m=1}^M \delta W_{im}^A G(y_m(k)) \\ &+ \sum_{k=0}^{N_g-1} \sum_{m=1}^M p_m^Z(k) G'(y_m(k)) \cdot \left[\sum_{j=1}^n \delta W_{mj}^B x_j(k) + \delta W_m^D \right] \\ &+ \sum_{k=0}^{N_g-1} \sum_{i \in VD} J_i(k) \sum_{m=1}^M W_{im}^A \delta G(y_m(k)) \\ &- \sum_{k=0}^{N_g-1} \sum_{i \in HD} p_i^X(k) [\delta x_i(k+1) - \sum_{m=1}^M W_{im}^A \delta G(y_m(k))] \\ &- \sum_{k=0}^{N_g-1} \sum_{m=1}^M p_m^Z(k) [\delta G(y_m(k)) - G'(y_m(k)) \cdot \sum_{j=1}^n W_{mj}^B \delta x_j(k)]. \end{aligned} \quad (A.7)$$

ここで、

$$p_i^X(k) = J_i(k) = \sum_{m=1}^M W_{im}^A G(y_m(k)) - q_i(k+1) \quad (i \in VD) \quad (A.8)$$

また、

$$p_m^Z(k-1) = \sum_{m=1}^M p_m^Z(k) G'(y_m(k)) W_{mi}^B, \quad (A.9)$$

$$p_i^X(N_g-1) = 0 \quad (i \in HD), \quad (A.10)$$

$$p_m^z(k) = \sum_{j=1}^n p_j^x(k) W_{jm}^A \quad (m = 1, \dots, M) \quad (\text{A.11})$$

とおき、

$$\delta x_i(k) \equiv 0 \quad (i \in VD)$$

の関係を利用すると、式(A.7)は次のようになる。

$$\begin{aligned} \delta E_1(\mathbf{w}) &= \sum_{k=0}^{N_g-1} \sum_{i=1}^n p_i^x(k) \sum_{m=1}^M \delta W_{im}^A G(y_m(k)) \\ &+ \sum_{k=0}^{N_g-1} \sum_{m=1}^M p_m^z(k) G'(y_m(k)) \cdot \left[\sum_{j=1}^n \delta W_{mj}^B x_j(k) + \delta W_m^D \right] \\ &+ \sum_{i \in HD} p_i^x(-1) \delta x_i(0). \end{aligned} \quad (\text{A.12})$$

よって、

$$\frac{\partial E_1(\mathbf{w})}{\partial W_{im}^A} = \sum_{k=0}^{N_g-1} p_i^x(k) G(y_m(k)). \quad (\text{A.13})$$

$$\frac{\partial E_1(\mathbf{w})}{\partial W_{mj}^B} = \sum_{k=0}^{N_g-1} p_m^z(k) G'(y_m(k)) x_j(k). \quad (\text{A.14})$$

$$\frac{\partial E_1(\mathbf{w})}{\partial W_m^D} = \sum_{k=0}^{N_g-1} p_m^z(k) G'(y_m(k)), \quad (\text{A.15})$$

$$(i = 1, \dots, n, \quad m = 1, \dots, M),$$

$$\frac{\partial E_1(\mathbf{w})}{\partial x_i(0)} = p_i^x(-1) \quad (i \in HD) \quad (\text{A.16})$$

となる。

A.1.2 軌道誤差法

$$J_i(k) = x_i(k) - q_i(k) \quad (i \in VD) \quad (\text{A.17})$$

とする。\$J_i(0) = 0\$であることを考慮すると、誤差評価関数 \$E_2\$ は次のように表現できる。

$$\begin{aligned} E_2(\mathbf{w}) &= \sum_{k=0}^{N_g} \left\{ \frac{1}{2} \sum_{i \in VD} J_i(k)^2 - \sum_{i \in HD} p_i^x(k) [x_i(k+1) - \sum_{m=1}^M W_{im}^A G(y_m(k))] \right. \\ &\quad \left. - \sum_{m=1}^M p_m^z(k) [G(y_m(k)) - G(\sum_{j=1}^n W_{mj}^B x_j(k) + W_m^D)] \right\}. \end{aligned} \quad (\text{A.18})$$

そこで、関数誤差法の場合と同様に、変分 δE_2 を算出し、

$$p_i^X(k-1) = \sum_{m=1}^M p_m^Z(k) G'(y_m(k)) W_{mi}^B + \delta_{iV} J_i(k). \quad (\text{A.19})$$

$$p_i^X(N_B) = 0 \quad (i = 1, \dots, n). \quad (\text{A.20})$$

とおく。ただし、

$$\delta_{iV} = \begin{cases} 1 & (i \in VD) \\ 0 & (i \in HD) \end{cases}$$

である。このとき、 δE_2 は式 (A.12) に一致する。よって、パラメータおよび初期値の更新則も関数誤差法の場合の式 (A.13) ~ (A.16) において E_1 を E_2 に変えたものに一致する。

A.2 連続力学系の学習則

速度誤差法と波形誤差法の誤差評価関数は、それぞれ次のように与えられる。

$$E_3(\omega) = \frac{1}{2} \int_0^{T_B} \|\dot{f}(\mathbf{x}(t), \omega) - f(\mathbf{q}(t))\|^2 dt, \quad (\text{A.21})$$

$$E_4(\omega) = \frac{1}{2} \int_0^{T_B} \|\mathbf{x}(t) - \mathbf{q}(t)\|^2 dt. \quad (\text{A.22})$$

ただし、速度誤差法において $\mathbf{x}(t)$ は

$$x_i(t) = q_i(t) \quad (i \in VD, t \in [0, T_B])$$

のように VD の成分を教師信号に拘束した RNN の出力であり、波形誤差法において $\mathbf{x}(k)$ は VD の初期値のみを

$$x_i(0) = q_i(0) \quad (i \in VD)$$

のように教師信号に一致させた RNN の出力波形である。

離散系の学習と同じく、次の勾配降下法に基づき、パラメータ ω と $x_i(0)$ を更新することにより学習を行なう。

$$\omega \leftarrow \omega - \eta \nabla_{\omega} E_l \quad (l = 3, 4). \quad (\text{A.23})$$

$$x_i(0) \leftarrow x_i(0) - \eta \frac{\partial E_l}{\partial x_i(0)} \quad (l = 3, 4, i \in HD). \quad (\text{A.24})$$

A.2.1 速度誤差法

まず、

$$J_i(t) = \dot{f}_i(\mathbf{x}(t), \mathbf{w}) - f_i(\mathbf{q}(t)) \quad (i \in VD) \quad (\text{A.25})$$

とし、誤差評価関数 E_3 を次のように表現する。

$$\begin{aligned} E_3(\mathbf{w}) = & \int_0^{T_B} \left\{ \frac{1}{2} \sum_{i \in VD} J_i(t)^2 - \sum_{i \in HD} p_i^X(t) \left[\frac{dx_i(t)}{dt} - \sum_{m=1}^M W_{im}^A G(y_m(t)) \right] \right. \\ & \left. - \sum_{m=1}^M p_m^Z(t) [G(y_m(t)) - G(\sum_{j=1}^n W_{mj}^B x_j(t) + W_m^D)] \right\} dt. \end{aligned} \quad (\text{A.26})$$

パラメータ \mathbf{w} の修正に対する E_3 の変分は次のように表現できる。

$$\begin{aligned} \delta E_3(\mathbf{w}) = & \int_0^{T_B} \sum_{i \in VD} J_i(t) \sum_{m=1}^M \delta W_{im}^A G(y_m(t)) dt + \int_0^{T_B} \sum_{i \in HD} p_i^X(t) \sum_{m=1}^M \delta W_{im}^A G(y_m(t)) dt \\ & + \int_0^{T_B} \sum_{m=1}^M p_m^Z(t) G'(y_m(t)) \cdot \left[\sum_{j=1}^n \delta W_{mj}^B x_j(t) + \delta W_m^D \right] dt \\ & + \int_0^{T_B} \sum_{i \in VD} J_i(t) \sum_{m=1}^M W_{im}^A \delta G(y_m(t)) dt \\ & - \int_0^{T_B} \sum_{i \in HD} p_i^X(t) \left[\delta \frac{dx_i(t)}{dt} - \sum_{m=1}^M W_{im}^A \delta G(y_m(t)) \right] dt \\ & - \int_0^{T_B} \sum_{m=1}^M p_m^Z(t) [\delta G(y_m(t)) - G'(y_m(t)) \cdot \sum_{j=1}^n W_{mj}^B \delta x_j(t)] dt. \end{aligned} \quad (\text{A.27})$$

ここで、

$$p_i^X(t) = J_i(t) = \sum_{m=1}^M W_{im}^A G(y_m(t)) - f_i(\mathbf{q}(t)) \quad (i \in VD). \quad (\text{A.28})$$

また、

$$\frac{dp_i^X(t)}{dt} = - \sum_{m=1}^M p_m^Z(t) G'(y_m(t)) W_{mi}^B, \quad (\text{A.29})$$

$$p_i^X(T_B) = 0 \quad (i \in HD), \quad (\text{A.30})$$

$$p_m^Z(t) = \sum_{j=1}^n p_j^X(t) W_{jm}^A \quad (m = 1, \dots, M) \quad (\text{A.31})$$

とおき、

$$\delta x_j(t) \equiv 0 \quad (j \in VD)$$

の関係を利用すると、式(A.27)は次のようになる。

$$\begin{aligned} \delta E_3(w) &= \int_0^{T_B} \sum_{i=1}^n p_i^X(t) \sum_{m=1}^M \delta W_{im}^A G(y_m(t)) dt \\ &+ \int_0^{T_B} \sum_{m=1}^M p_m^Z(t) G'(y_m(t)) \cdot \left[\sum_{j=1}^n \delta W_{mj}^B x_j(t) + \delta W_m^D \right] dt \\ &+ \sum_{i \in HD} p_i^X(0) \delta x_i(0) \end{aligned} \quad (\text{A.32})$$

よって、

$$\frac{\partial E_3(w)}{\partial W_{im}^A} = \int_0^{T_B} p x_i(t) G(y_m(t)) dt, \quad (\text{A.33})$$

$$\frac{\partial E_3(w)}{\partial W_{mi}^B} = \int_0^{T_B} p z_m(t) G'(y_m(t)) x_i(t) dt, \quad (\text{A.34})$$

$$\frac{\partial E_3(w)}{\partial W_m^D} = \int_0^{T_B} p z_m(t) G'(y_m(t)) dt, \quad (\text{A.35})$$

$$(i = 1, \dots, n, \quad m = 1, \dots, M),$$

$$\frac{\partial E_3(w)}{\partial x_i(0)} = p_i^X(0) \quad (i \in HD) \quad (\text{A.36})$$

となる。

A.2.2 波形誤差法

まず、

$$J_i(t) = x_i(t) - q_i(t) \quad (i \in VD) \quad (\text{A.37})$$

とおくと、誤差評価関数 E_4 は速度誤差法の場合の表現式(A.26)と一致する。

そこで、変分 δE_4 を算出し、

$$\frac{dp_i^X(t)}{dt} = - \sum_{m=1}^M p_m^Z(t) G'(y_m(t)) W_{mi}^B - \delta_{iV} J_i(t), \quad (\text{A.38})$$

$$p_i^X(T_B) = 0 \quad (i = 1, \dots, n) \quad (\text{A.39})$$

とおけば、式(A.33)～(A.36)において E_3 を E_4 に変えたものが波形誤差法の更新則になる。

付録 B リアブノフ指数と法線リアブノフ指数

次のような $n+1$ 次元離散力学系を考える。

$$\begin{pmatrix} w(k+1) \\ q(k+1) \end{pmatrix} = F \begin{pmatrix} w(k) \\ q(k) \end{pmatrix} = \begin{pmatrix} g(w(k), q(k)) \\ f(q(k)) \end{pmatrix} \quad (w \in \mathbb{R}^n, q \in \mathbb{R}) \quad (\text{B.1})$$

関数 F のヤコビ行列は、 f が w に依存しないことから次のように書けることがわかる。

$$DF \begin{pmatrix} w \\ q \end{pmatrix} = \left(\begin{array}{c|c} & \partial g_1 / \partial q \\ \hline D_w g(w, q) & \cdot \\ & \cdot \\ & \cdot \\ \hline \mathbf{0} & \partial f / \partial q \end{array} \right)$$

ここで、 f に関するリアブノフ指数

$$\lambda_{n+1} = \lim_{k \rightarrow \infty} \frac{1}{k} \log \left| \frac{df^k(q(0))}{dq} \right| \quad (\text{B.2})$$

と、 n 個の法線リアブノフ指数

$$\lambda_j = \lim_{k \rightarrow \infty} \frac{1}{k} \log \| D_w g^{(k)}(w(0), q(0)) w_j \| \quad (1 \leq j \leq n) \quad (\text{B.3})$$

が存在するならば、系 (B.1) のリアブノフ指数は、 $\lambda_1, \dots, \lambda_{n+1}$ になる。ただし、 $D_w g^{(k)}(w(0), q(0))$ は $DF^k(w(0), q(0))$ の左上 $n \times n$ 成分の行列である。

(証明) 式 (B.3) で表わされた n 個の法線リアブノフ指数を与えるベクトルを w_j とし、 $v_j = (w_j, 0)^T$ ($j = 1, \dots, n$) とすると、

$$\lim_{k \rightarrow \infty} \frac{1}{k} \log \| D F^k(w(0), q(0)) v_j \| = \lim_{k \rightarrow \infty} \frac{1}{k} \log \| D_w g^{(k)}(w(0), q(0)) w_j \| = \lambda_j$$

となるから、 $\lambda_1, \dots, \lambda_n$ は確かに系 (B.1) のリアブノフ指数である。

また、 $k \geq 1$ について、 $f^k(q(0))$ は w に依存しないから、

$$DF^k \begin{pmatrix} w \\ q \end{pmatrix} = \left(\begin{array}{c|c} D_{wg^{(k)}}(w, q) & \begin{matrix} * \\ * \\ * \end{matrix} \\ \hline 0 & \partial f^k / \partial q \end{array} \right)$$

と書ける。よって、

$$DF^k \begin{pmatrix} w \\ q \end{pmatrix} \text{ の固有値の積} = \frac{\partial f^k}{\partial q} \times D_{wg^{(k)}}(w, q) \text{ の固有値の積}$$

となる。よって、系(B.1) のリアプノフ指数を $\lambda_1, \dots, \lambda_n, \lambda'_{n+1}$ とすると、

$$\begin{aligned} \lambda_1 + \dots + \lambda_n + \lambda'_{n+1} &= \lim_{k \rightarrow \infty} \frac{1}{k} \log \sqrt{DF^k \begin{pmatrix} w \\ q \end{pmatrix} \cdot [DF^k \begin{pmatrix} w \\ q \end{pmatrix}]^t \text{ の固有値の積}} \\ &= \lim_{k \rightarrow \infty} \frac{1}{k} \log \left| \frac{\partial f^k}{\partial q} \right| + \lim_{k \rightarrow \infty} \frac{1}{k} \log \sqrt{D_{wg^{(k)}}(w, q) \cdot [D_{wg^{(k)}}(w, q)]^t \text{ の固有値の積}} \\ &= \lambda_1 + \dots + \lambda_n + \lambda_{n+1} \end{aligned}$$

となる。よって、

$$\lambda'_{n+1} = \lambda_{n+1}$$

である。

□

付録 C 法線リアブノフ指数と引力圏の測度に関する考察

第6章の定理6-1の証明において、「2次元離散力学系の1次元閉不変集合 A の法線リアブノフ指数が負であれば、 A の引力圏の測度は正である⁴⁶⁾。」という命題を利用した。ここでは、力学系を記述する関数が局所的に線形な場合について、これを証明し、文献(48)の議論を補足する。また、より一般的な場合について若干の考察を加える。

まず、問題の設定といくつかの記号の定義を行なう。

次のような2次元離散力学系を考える。

$$\begin{pmatrix} x(k+1) \\ y(k+1) \end{pmatrix} = F \begin{pmatrix} x(k) \\ y(k) \end{pmatrix} = \begin{pmatrix} f_1(x(k)) \\ f_2(x(k), y(k)) \end{pmatrix} \quad (x \in \mathbf{R}, y \in \mathbf{R}). \quad (\text{C.1})$$

ここで、1次元離散系

$$x(k+1) = f_1(x(k))$$

は不変集合 A を持ち、その不変確率測度を μ とする。

正数 E に対して、

$$\begin{aligned} M_E &= A \times \{ y \in \mathbf{R} \mid |y| < E \}, \\ U_E &= \{ (x, y) \in \mathbf{R}^2 \mid F^k(x, y) \in M_E \ (\forall k \geq 0) \text{ and } \omega((x, y)) \subset A \} \end{aligned}$$

とする。ここで、 $\omega((x, y))$ は点 (x, y) の ω 極限集合を表わす。

また、1次元ルベーク測度を l で表わし、 \mathbf{R}^2 上の測度 $\hat{\mu}$ を $\hat{\mu} = \mu \times l$ で定義する。さらに $\epsilon < E$ となる正数 ϵ に対して、2次元可測集合 M の ϵ -相対測度 $\hat{\mu}_\epsilon$ を

$$\hat{\mu}_\epsilon(M) = \frac{\hat{\mu}(M \cap M_\epsilon)}{\hat{\mu}(M_\epsilon)}$$

で定義する。

また、不変集合 A の法線リアブノフ指数を

$$\lambda = \lim_{k \rightarrow \infty} \frac{1}{k} \log \|D_x f_2^{(k)}(x, 0)\| \quad (\mu\text{-a.e. } x) \quad (\text{C.2})$$

とする。ただし、 $D_x f_2^{(k)}(x, y)$ は $DF^k(x, y)$ の(2,2)成分である。

このとき、次の定理が成立する。

[定理 C-1] $f_2(x, y)$ は $y = 0$ の E 近傍で y に関して線形であるとする。すなわち、 $|y| < E$ ならば、

$$f_2(x, y) = D_y f_2(x, 0)y \quad (C.3)$$

であるとする。

このとき、法線リアプノフ指数が $\lambda < 0$ であれば、

$$\lim_{\rightarrow 0} \hat{\mu}_\epsilon(U_E) = 1$$

となる。

□

A の引力圏を $B(A)$ とすると、 $B(A) \supset U_E$ である。また、定理 C-1 より、 f_2 が y に関して局所的に線形で、 A の法線リアプノフ指数が負であれば、 $\lim_{\epsilon \rightarrow 0} \hat{\mu}_\epsilon(U_E) = 1$ であるから、 $\hat{\mu}(U_E) > 0$ である。したがって、 A の不変測度 μ が 1 次元ルベーク測度に対して絶対連続であれば、 $B(A)$ の 2 次元ルベーク測度は正になる。

定理の証明の前に、補題を一つ述べる。

まず、 $x \in A$ に対して関数 $\tau(x)$ を次のように定義する。

$$\tau(x) = \sup_{k \geq 0} \left| \exp\left(-\frac{3}{4}k\lambda\right) D_y f_2^{(k)}(x, 0) \right|.$$

式 (C.2) の極限は μ に関してほとんどすべての x について同一の値 λ になり、かつ $\lambda < 0$ であるから、ほとんどすべての x について $\tau(x)$ は有限の値となる。したがって、

$$R(s) = \{ x \in A \mid \tau(x) \leq s \}$$

とすると、

$$\lim_{s \rightarrow \infty} \mu(R(s)) = 1$$

である。

[補題 C-1] $s \geq 1$ で $x \in R(s)$ かつ $|y| < E/s$ ならば、任意の整数 $k \geq 0$ に対して、

$$|f_2^{(k)}(x, y)| \leq \exp\left(-\frac{1}{2}k\lambda\right) |D_y f_2^{(k)}(x, 0)| \cdot |y|$$

である。

(証明) k に関する帰納法。

(i) $k=0$ のとき。

$f_2^{(0)}(x, y) = y$, $D_y f_2^{(0)}(x, 0) = 1$, $\exp(-1/2 \cdot 0 \cdot \lambda) = 1$ より明らかに成立。

(ii) $k=K-1$ で成立しているとして、 $k=K$ のとき。

まず、 $x \in R(s)$ だから $(x, 0) \in A$ であり、かつ、 A は f_1 の不変集合だから、

$$(f_1^{(K-1)}(x), 0) \in A.$$

また、帰納法の仮定より、

$$|f_2^{(K-1)}(x, y)| \leq \exp(-\frac{1}{2}(K-1)\lambda) |D_y f_2^{(K-1)}(x, 0)| \cdot |y|.$$

ここで、 $x \in R(s)$ だから、

$$|D_y f_2^{(K-1)}(x, 0)| \leq s \cdot \exp(\frac{3}{4}(K-1)\lambda).$$

したがって、

$$|f_2^{(K-1)}(x, y)| \leq \exp(\frac{1}{4}(K-1)\lambda) s \cdot E/s < E.$$

よって、 $(f_1^{(K-1)}(x), f_2^{(K-1)}(x, y)) \in M_E$ である。したがって、局所線形性の仮定 (C.3) より、

$$\begin{aligned} |f_2^{(K)}(x, y)| &= |f_2(f_1^{(K-1)}(x), f_2^{(K-1)}(x, y))| \\ &\leq \exp(-\frac{1}{2}\lambda) |D_y f_2(f_1^{(K-1)}(x), 0)| \cdot |f_2^{(K-1)}(x, y)|. \end{aligned}$$

そこで、再び帰納法の仮定より、

$$\begin{aligned} |f_2^{(K)}(x, y)| &\leq \exp(-\frac{1}{2}\lambda) |D_y f_2(f_1^{(K-1)}(x), 0)| \cdot \exp(-\frac{1}{2}(K-1)\lambda) \cdot |D_y f_2^{(K-1)}(x, 0)| \cdot |y| \\ &= \exp(-\frac{1}{2}K\lambda) |D_y f_2^{(K)}(x, 0)| \cdot |y| \end{aligned}$$

となって、 $k=K$ でも成立する。 □

(定理 C-1 の証明) $x \in R(s)$ かつ $|y| < E/s$ ならば補題 C-1 と $R(s)$ の定義より、

$$|f_2^{(k)}(x, y)| \leq s \exp(\frac{1}{4}k\lambda) |y|$$

が任意の $k \geq 0$ について成り立つ。ただし、 $s \geq 1$ とする。よって、 $x \in R(s)$ かつ $|y| < E/s$ ならば、任意の $k \geq 0$ について、

$$F^k(x, y) \in M_E$$

となり、かつ

$$\lim_{\epsilon \rightarrow \infty} f_2^{(k)}(x, y) = 0$$

である。よって、 $\epsilon \leq E/s$ とすると、 $U_\epsilon \supset R(s) \times \{y \mid |y| < \epsilon\}$ だから、

$$\begin{aligned} \hat{\mu}_\epsilon(U_\epsilon) &= \frac{\hat{\mu}(U_\epsilon \cap M_\epsilon)}{\hat{\mu}(M_\epsilon)} \geq \frac{\hat{\mu}(R(s) \times \{y \mid |y| < \epsilon\})}{\hat{\mu}(A \times \{y \mid |y| < \epsilon\})} \\ &= \mu(R(s))/\mu(A) = \mu(R(s)). \end{aligned}$$

そこで、 $s = E/\epsilon$ とすると、

$$\lim_{\epsilon \rightarrow 0} \hat{\mu}_\epsilon(U_\epsilon) = \lim_{s \rightarrow \infty} \mu(R(s)) = 1$$

となる。

□

また、局所線形性を仮定しない場合は、次のような結果が一般の多次元系の場合に成立する。

ここでは、系(C.1)の状態変数 x, y をそれぞれ、 m 次元、 n 次元ベクトルとする。また、正数 E に対して、

$$\begin{aligned} M_E &= A \times \{y \mid \|y\| < E\}, \\ V_E &= \{(x, y) \mid (x, y) \in M_E \text{ and } \omega((x, y)) \subset A\} \end{aligned}$$

とする。

A の不変確率測度を μ 、 n 次元ルベーク測度を l_n として、 R^{m+n} 上の測度 $\hat{\mu}$ を $\hat{\mu} = \mu \times l_n$ で定義する。また、

$$\hat{\mu}(V_E) = \frac{\hat{\mu}(V_E)}{\hat{\mu}(M_E)}$$

とする。このとき、次の定理が成り立つ。

[定理C-2] $x \in A$ について関数 $\delta(x)$ を

$$\delta(x) = \sup_{\delta \geq 0} \{ \delta \in R \mid \|y\| \leq \delta \Rightarrow \omega((x, y)) \subset A \}$$

と定義する ($0 \leq \delta(x) \leq +\infty$)。

このとき、 A が条件

$$\mu(\{x \in A \mid \delta(x) > 0\}) = 1 \tag{C.4}$$

* V_E は U_E よりも条件が弱い

を満たしているならば、

(1) $E > 0$ が存在して、 $\hat{\mu}(V_E) > 0$.

(2) $\lim_{E \rightarrow 0} \hat{\mu}(V_E) = 1$.

(証明) (1) は (2) から帰結するが、個別に証明を与えておく。

((1) の証明) まず、集合 A_E を

$$A_E = \{ x \in A \mid \delta(x) > E \}$$

と定義する。もし、任意の $E > 0$ について $\hat{\mu}(V_E) = 0$ とすると、任意の $E > 0$ について $\mu(A_E) = 0$ となる。したがって、

$$A_0 = \bigcup_{k=1}^{\infty} A_{\frac{1}{k}}$$

より、 $\mu(A_0) = 0$ となって、条件 (C.4) に矛盾する。

((2) の証明) $\lim_{E \rightarrow 0} \hat{\mu}(V_E) \neq 1$ とすると、ある $\alpha < 1$ について、任意の $\epsilon > 0$ に対し、 $E(\epsilon) < \epsilon$ となる $E(\epsilon)$ が存在して、 $\hat{\mu}(V_{E(\epsilon)}) \leq \alpha$ となる。このとき、 $\mu(A_{2E(\epsilon)}) \leq \alpha$ である。したがって、

$$A_0 = \bigcup_{k=1}^{\infty} A_{2E(\frac{1}{k})}$$

であり、 $k_1 > k_2$ ならば、 $A_{2E(\frac{1}{k_1})} \supset A_{2E(\frac{1}{k_2})}$ とできるから、 $\mu(A_0) \leq \alpha < 1$ となって、条件 (C.4) に矛盾する。 \square

条件 (C.4) は A の最大法線リアブノフ指数が負であれば一般に成立するように思われる⁴⁵⁾。

参考文献

- 1) ブリゴジン, I. : "存在から発展へ" (小出, 我孫子訳), みすず書房 (1984)
- 2) Guckenheimer, J. and Holmes, P. : "Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields", Springer-Verlag (1983)
- 3) Eckmann, J. P. and Ruelle, D. : "Ergodic Theory of Chaos and Strange Attractors", *Reviews of Modern Physics*, 57, 3, pp. 617-656 (1985)
- 4) Devaney, R. L. : "An Introduction to Chaotic Dynamical Systems", Benjamin/Cummings Publishing, Inc. (1986)
- 5) Thompson J. M. T. and Stewart H. B. : "Nonlinear Dynamics and Chaos", John Wiley & Sons (1986)
- 6) 高安秀樹編 : "フラクタル科学" (第3章), 朝倉書店 (1987)
- 7) ザスラフスキー, G. M. : "カオス" - 古典および量子力学系 (三島, 齊藤, 新藤訳), 現代工学社 (1989)
- 8) 合原一幸編 : "カオス" - カオス理論の基礎と応用, サイエンス社 (1990)
- 9) 長島弘幸, 馬場良和 : "カオス入門" - 現象の解析と数理, 培風館 (1992)
- 10) Hale, J. K. : "Ordinary Differential Equations", John Wiley & Sons (1969)
- 11) Willems, J. L. : "Stability Theory of Dynamical Systems", Thomas Nelson & Sons (1970)
- 12) 斎藤利弥 : "位相力学", 共立出版 (1971)
- 13) 杉山昌平 : "差分・微分方程式", 共立出版 (1971)
- 14) 白岩謙一 : "力学系の理論", 岩波書店 (1974)
- 15) 青木統夫, 白岩謙一 : "力学系とエントロピー", 共立出版 (1985)
- 16) Farmer, J. D. and Sidorowich, J. J. : "Predicting Chaotic Time Series", *Phys. Rev. Lett.*, 59, 8, pp. 845-848 (1987)

- 17) Casdagli, M. : "Nonlinear Prediction of Chaotic Time Series". *Physica* . **35D**, pp. 335-356 (1989)
- 18) Sugihara, G. and May, M. M. : "Nonlinear Forecasting as a Way of Distinguishing Chaos from Measurement Error in Time Series". *Nature*. **344**, pp. 734-741 (1990)
- 19) Lapedes, A. and Farber, R. : "How Neural Nets Work". in *Neural Information Processing Systems*. Anderson, D. Z. ed., AIP Press (1988)
- 20) Funahashi, K. : "On the Approximate Realization of Continuous Mappings by Neural Networks", *Neural Networks*. **2**, pp. 183-192 (1989)
- 21) 安達雅春, 合原一幸, 小谷誠 : "バックプロパゲーション・ニューラルネットワークによる決定論的非線形予測", 信学技報, **NLP91-30** (1991)
- 22) Adachi, M. and Kotani, M. : "Identification of Chaotic Dynamical Systems with Back Propagation Neural Networks", *IEICE Trans. Fundamentals*, **E-77A**, 1, pp. 324-334 (1994)
- 23) 村上由彦, 佐藤雅昭 : "リカレントネットによる非線形ダイナミクスの近似とカオスの学習", 信学技報, **NC90-81** (1990)
- 24) Sato, M., Joe, K. and Hirahara, T. : "APOLONN brings us to the real world: Learning nonlinear dynamics and fluctuations in nature". *Proc. of IJCNN90*, **1**, pp. 581-587 (1990)
- 25) Sato, M. : "A Learning Algorithm to Teach Spatiotemporal Patterns to Recurrent Neural Networks", *Biol. Cybern.*, **62**, pp. 259-263 (1990)
- 26) Pearlmutter, B.A. : "Learning State Space Trajectories in Recurrent Neural Networks", *Neural Computation*, **1**, pp. 263-269 (1989)
- 27) Williams, R. J. and Zipser, D. : "A Learning Algorithm for Continually Running Fully Recurrent Neural Networks", *Neural Computation*, **1**, pp. 270-280 (1989)
- 28) 網谷賢治, 吉澤修治 : "時間パターンの記憶の神経回路網モデル", 信学論(D-II), **J73-D-II**, 8, pp. 1150-1157 (1990)
- 29) 浦浜喜一 : "ニューラルネットによる時間波形の適応学習", *Proc. of The First NOLTA Workshop*, pp. 27-35 (1991)

- 30) 松業育雄：“バックプロパゲーションによる特徴抽出”，数理科学，No. 338, pp. 31-37 (1991)
- 31) 渡辺辰巳，郷原一寿，内川嘉樹：“リカレントニューラルネットワークの各学習測に関する検討および学習曲面の形状”，信学論(D-II), J74-D-II, No. 12, pp. 1776-1787 (1991)
- 32) 本堂毅：“カオス時系列の神経回路網による学習過程”，信学技報，NLP92-38 (1992)
- 33) Hondou, T. and Sawada, Y.: “Dynamical Properties of Learning Processes of Chaotic Time Series by a Neural Network”, Proc. of NOLTA93, 1, pp. 217-222 (1993)
- 34) 黒江康明：“リカレントニューラルネットワークの学習法”，システム／制御／情報，36, 10, pp. 634-643 (1992)
- 35) 甘利俊一：“神経回路網の数理”- 脳の情報処理様式，産業図書(1978)
- 36) Rumelhart, D. E., McClelland, J. L. and the PDP research group eds.: “*Parallel Distributed Processing*”, MIT Press (1986)
- 37) 麻生英樹：“ニューラルネットワーク情報処理”，産業図書(1988)
- 38) 合原一幸：“ニューラルコンピュータ”，東京電機大学出版会(1988)
- 39) Hecht-Nielsen R.: “*Neurocomputing*”, Addison Wesley (1990)
- 40) 松橋誠壽：“ニューロコンピューティング入門”，オーム社(1992)
- 41) 合原一幸編：“ニューラルシステムにおけるカオス”，東京電機大学出版局(1993)
- 42) Rumelhart, D. E., Hinton, G. E. and Williams, R. J.: “Learning Internal Representations by Error Propagation”, in *Parallel Distributed Processing*. Vol. 1, Rumelhart, D. E., McClelland, J. L. and the PDP research group eds., MIT Press (1986)
- 43) Amari, S.: “A Theory of Adaptive Pattern Classifiers”, *IEEE Trans. Electronic Computers*, EC-16, No. 3, pp. 299-307 (1967)
- 44) 甘利俊一：“情報理論II”，共立出版(1968)

- 45) Lyapunov, A. M. : "The General Problem of the Stability of Motion", Taylor & Francis (1992)
- 46) Carroll, T. L. and Pecora, L. M. : "Synchronizing Chaotic Circuits". *IEEE Trans. Circuits Syst., CAS-38*, No. 4, pp. 453-456 (1991)
- 47) Endo, T. and Chua, L. O. : "Synchronizing Chaos from Electronic Phase-Locked Loops". *International Journal of Bifurcation and Chaos*, 1, No. 3, pp. 701-710 (1991)
- 48) Alexander, J. C., Yorke, J. A., You, Z. and Kan, I. : "Riddled Basins". *International Journal of Bifurcation and Chaos*, 2, No. 4, pp. 795-813 (1992)
- 49) Ott, E., Alexander, J. C., Kan, I., Sommerer, J. C. and Yorke, J. A. : "The Transition to Chaotic Attractors with Riddled Basins", *Physica D76*, pp. 384-410 (1994)
- 50) Pikovsky, A. S. and Grassberger, P. : "Symmetry Breaking Bifurcation for Coupled Chaotic Attractors", *J. Phys. A*, 24, pp. 4587-4597 (1991)
- 51) Ashwin, P., Buescu, J. and Stewart, I. : "Bubbling of Attractors and Synchronisation of Chaotic Oscillators", *Phys. Lett. A*, 193, pp. 126-139 (1994)
- 52) Heagy, J. F., Carroll, T. L. and Pecora, L. M. : "Experimental and Numerical Evidence for Riddled Basins in Coupled Chaotic Systems", *Phys. Rev. Lett.*, 73, 26, pp. 3528-3531 (1994)
- 53) Rosenblatt, F. : "The Perceptron : A Probabilistic Model for Information Storage and Organization in the Brain", *Psychol. Rev.*, 65, 6, pp. 386-408 (1958)
- 54) Minsky, M. L. and Papert, S. A. : "Perceptrons", MIT Press (1969)
- 55) Billingsley, P. : "Ergodic Theory and Information", John Wiley & Sons, New York (1960)
- 56) 今野 浩, 山下 浩 : "非線形計画法", 日科技連 (1978)
- 57) Oseledec, V. I. : "A Multiplicative Ergodic Theorem - Ljapunov Characteristic Numbers for Dynamical Systems", *Trans. Moscow Math. Soc.*, 19, pp. 197-231 (1968)
- 58) Shimada, I. and Nagashima, T. : "A Numerical Approach to Ergodic Problem of Dissipative Dynamical Systems", *Progr. Theor. Phys.*, 61, 6, pp. 1605-1616 (1979)

- 59) 上田皖亮 : "非線形性に基づく確率統計現象", 電学論. A-98, No. 3, pp. 167-173 (1978)
- 60) Takens, F. : "Detecting Strange Attractors in Turbulence". in *Lecture Notes in Mathematics*. Springer-Verlag (1981)
- 61) Lorenz, E. N. : "The Local Structure of a Chaotic Attractor in Four Dimensions". *Physica D13*, pp. 90-104 (1984)
- 62) Milnor, J. : "On the Concept of Attractor". *Commun. Math. Phys.*, **99**, pp. 177-195 (1985)
- 63) Hondou, T. : "Symmetry Breaking by Correlated Noise in a Multistable System". *Journal of Physical Society of Japan*, **63**, 5, pp. 2014-2015 (1994)
- 64) Radons, G., Schuster, H. G. and Werner, D. : "Drift and Diffusion in Backpropagation Networks", in *Parallel Processing in Neural Systems and Computers*, Eckmiller, R. Hartmann, G. and Hauske, G. eds., North-Holland (1990)

本研究に関する著者の発表

- [101] 上田 皖亮, 中島 弘之 : “動揺時の電力系統を記述する非線形方程式の解についての一考察”, 電気関係学会関西支部連合大会講演論文集 G9 (1985)
- [102] 上田 皖亮, 中島 弘之, 引原 隆士 : “磁気弾性系における強制振動”, 電子通信学会技術研究報告 NLP85-38 (1985)
- [103] Ueda, Y., Nakajima, H., Hikihara, T. and Stewart, H. B. : “Forced Two-well Potential Duffing’s Oscillator.” in *Dynamical Systems Approaches to Nonlinear Problems in Systems and Circuits*, Salam, F. M. A. and Levi M. L. eds., pp. 128 -137, SIAM (1988)
- [104] 中島 弘之, 上田 皖亮 : “2 自由度強制振動系に発生するカオスに関する一考察”, 電気関係学会関西支部連合大会講演論文集 G14 (1990)
- [105] 中島 弘之, 上田 皖亮 : “カオスアトラクタの分岐と混合性に関する一考察”, 電子情報通信学会春季全国大会講演論文集, A-86 (1991)
- [106] 中島 弘之, 長尾 茂樹, 上田 皖亮 : “2 自由度強制振動系におけるカオスアトラクタの分岐”, 電子情報通信学会春季全国大会講演論文集, A-87 (1991)
- [107] Nakajima, H. and Ueda, Y. : “Bifurcation Phenomena in a Two-degrees-of-freedom Duffing’s Equation”, IEICE Transactions E-74, No. 6, pp. 1414-1419 (1991)
- [108] Nakajima, H. and Ueda, Y. : “A Computer Experiment on the Mixing Property and the Orbital Instability of Chaotic Attractors”, Proc. of The Second Symposium on Nonlinear Theory and its Applications, pp. 91-94 (1991)
- [109] 中島 弘之, 長尾 茂樹, 上田 皖亮 : “結合 Duffing 方程式系におけるクライシスについて”, 電子情報通信学会技術研究報告 NLP91-42 (1991)
- [110] 中島 弘之, 上田 皖亮 : “非混合的カオスの相互情報量について”, 電気関係学会関西支部連合大会講演論文集 G3 (1991)
- [111] 中島 弘之, 上田 皖亮 : “カオスアトラクタの局所構造分解の一方法”, 電子情報通信学会春季全国大会講演論文集, A-68 (1992)

- [112] Nakajima, H. and Ueda, Y.: "Lyapunov Exponents for Differential-difference Equations." Proc. of 1992 Symposium on Nonlinear Theory and its Applications, pp. 115-118 (1992)
- [113] 中島弘之, 上田 皖亮: "微分差分方程式のリアプノフ指数", 電気関係学会関西支部連合大会講演論文集 G4 (1992)
- [114] 中島弘之, 甲田 哲也, 上田 皖亮: "リカレントネットによる離散力学系の学習過程について", 電子情報通信学会技術研究報告, CAS92-92, NLP92-72 (1993)
- [115] 中島弘之, 甲田 哲也, 上田 皖亮: "リカレントネットによるカオスアトラクタの学習過程", 電子情報通信学会春季全国大会講演論文集, A-82 (1993)
- [116] Nakajima, H., Koda, T. and Ueda, Y.: "A Measure Theoretical Analysis of Learning Algorithms for Recurrent Neural Networks." Proc. of International Joint Conference on Neural Networks, NAGOYA, 3, pp. 2575-2578 (1993)
- [117] 中島弘之, 上田 皖亮: "リカレントニューラルネットワークにおけるカオス学習の加速について", 電気関係学会関西支部連合大会講演論文集 G1-30 (1993)
- [118] Nakajima, H. and Ueda, Y.: "Process of Learning Chaotic Attractors in Recurrent Neural Networks." Proc. of NOLTA93, 2, pp. 601-604 (1993)
- [119] 中島弘之, 甲田 哲也, 上田 皖亮: "リカレントニューラルネットワークによる離散力学系の学習過程", 電子情報通信学会論文誌, J77-A No. 1, pp. 24-31 (1994)
- [120] 中島弘之, 上田 皖亮: "バックプロパゲーション法による1次元離散力学系学習の収束について", 電子情報通信学会技術研究報告 NLP93-75 (1994)
- [121] 中島弘之, 上田 皖亮: "ニューラルネットワーク結合パラメータが示すカオスの挙動", 電子情報通信学会春季全国大会講演論文集, A-78 (1994)
- [122] 中島弘之, 大西隆史, 上田 皖亮: "ニューラルネットワークによる1次元離散力学系の学習における教師信号提示方法について", 電子情報通信学会技術研究報告 NLP94-13 (1994)
- [123] 中島弘之, 上田 皖亮: "連続力学系の学習過程の非自律的差分方程式の理論に基づく解析", 電子情報通信学会技術研究報告 NLP94-48 (1994)

- [124] Nakajima, H. and Ueda, Y.: "An Analysis of the Back-Propagation Learning of Discrete Dynamical Systems Based on the Theory of Non-autonomous Difference Equation." Proc. of NOLTA94, pp. 125-128 (1994)
- [125] 中島弘之, 上田 皖亮: "リカレントニューラルネットワークによる連続力学系の学習の収束過程の解析", 電気関係学会関西支部連合大会講演論文集 G1-30 (1994)
- [126] 中島弘之, 上田 皖亮: "バックプロパゲーション法による離散力学系の学習の非自律的差分方程式の理論を用いた解析", 電子情報通信学会論文誌, J77-D-II No. 11, pp. 2279-2286 (1994)
- [127] 藤井丈仁, 中島弘之, 上田 皖亮: "不安定化制御を用いた周期軌道の学習", 電子情報通信学会総合大会講演論文集, A-75 (1995)
- [128] 中島弘之, 上田 皖亮: "勾配降下法によるカオスの学習に現れる Riddled Basin", 電子情報通信学会総合大会講演論文集, SA-2-12 (1995)
- [129] 中島弘之, 上田 皖亮: "力学系の学習における最適点の引力圏について", 電子情報通信学会技術研究報告 NLP95-7 (1995)
- [130] 中島弘之, 上田 皖亮: "Pyragas の方法によるカオス制御におけるゲインの自動調整", 電子情報通信学会技術研究報告 NLP95-32 (1995)
- [131] 中島弘之, 上田 皖亮: "Pyragas の DFC 法によるカオス制御におけるゲインの自動調整", 電気関係学会関西支部連合大会講演論文集 G1-3 (1995)
- [132] 中島弘之, 藤井丈仁, 上田 皖亮: "不安定化制御を用いた 2 次元離散系の周期軌道の学習", 電気関係学会関西支部連合大会講演論文集 G1-4 (1995)
- [133] 中島弘之, 上田 皖亮: "カオス同期における Riddled Basin のモデル", 電子情報通信学会技術研究報告 NLP95-66 (1995)
- [134] 中島弘之, 上田 皖亮: "ニューラルネットワークによる連続力学系の学習過程の非自律的差分方程式の理論に基づく解析", 電子情報通信学会論文誌, J78-D-II No. 12, pp. 1890-1898 (1995)