NAIST-IS-DT0061025

博士論文

部分観測システム同定の応用と脳型情報処理に関する研究

吉田 和子

2003 年 3 月 7 日

奈良先端科学技術大学院大学
情報科学研究科 情報システム学専攻

本論文は奈良先端科学技術大学院大学情報科学研究科に
学位 (理学) 授与の要件として提出した博士論文である。

吉田 和子

審査委員： 石井 信 教授
湊 小太郎 教授
杉本 謙二教授

部分観測システム同定の応用と脳型情報処理に関する研究*

吉田 和子

内容梗概

ヒトを取り巻く自然環境は、直接観測できない事象を含み、時間と共に動的に変化する。このような複雑な環境においても、ヒトは環境の特性を学習し、その変動を予測することによって最適な行動を取る。つまり、部分観測環境におけるシステム同定を行い、それに基づいて最適意思決定問題を解いていると考えられる。本研究では、部分観測環境におけるシステム同定問題について、機械学習と脳型学習の両側面から議論する。

まず初めに、統計的手法による非線形力学系のシステム同定問題について議論する。関数近似器として正規化ガウス関数ネットワークを用い、オンライン EM アルゴリズムによって学習を行う。実験では、低次元カオスシステムを取り上げ、力学変数の 1 変数の軌道の観測に基づき、元のカオス力学系のダイナミクスを学習させる。本研究では、部分観測データから元の状態空間での位相構造を構成する手法として、2 種類の埋め込み法について検討する。初めに、力学系の同定に一般に用いられる遅れ座標埋め込み手法を適用し、部分観測においてもシステムを同定できることを示す。次に、この手法を発展させた IIR(infinite impulse response) フィルタを用いた埋め込み法を提案する。ノイズを付加したデータを学習させることによって、学習時におけるノイズの影響を調べた結果、新しい埋め込み手法は従来のものと比べてさらにノイズに頑強であることを確認した。

次に、部分観測環境におけるシステム同定とそれに基づく意思決定法について議論する。機械学習において、最適性を報酬に基づき定義すると、確率的環境に

*奈良先端科学技術大学院大学 情報科学研究科 情報システム学専攻 博士論文, NAIST-IS-DT0061025, 2003 年 3 月 7 日.

おける最適意思決定問題は、マルコフ決定過程として定式化される。直接観測できない変数 (隠れ変数) が存在する環境は、部分観測マルコフ決定過程としてモデル化する。本研究では、隠れ変数をもつ環境でのモデル同定強化学習法を提案する。モデル同定強化学習は、過去の経験から環境を同定し、それに基づいて意思決定を行う。本手法では、環境の同定に忘却効果を導入したベイズ推定を用いる。また、動的に変化する環境を扱うため、環境の変化を検出しそれに基づいて行動様式を変化させる手法を提案する。本手法を隠れ変数を持つ迷路探索問題に適用した結果、従来の手法よりも環境の変化にうまく適応できることが分かった。

最後に、モデル同定強化学習を実現する脳の情報処理モデルを提案する。相互作用によって環境システムを同定し、それに基づいて意思決定を行うという学習過程は、ヒトの学習法としても妥当である。また、近年の研究で、いくつかの強化学習アルゴリズムが脳神経系のネットワークと関連付けられることが指摘されている。本研究では、モデル同定強化学習で用いる主な関数と、脳、特に前頭前野の機能を対応付けることにより、前頭前野強化学習モデルを提案する。このモデルでは、背外側前頭前野が報酬依存環境モデルの保持と操作に、前部前頭前野が観測不可能な環境の推定に関わるとみなす。また、これらの推定に基づいて行動選択を行う場として、帯状回を想定する。このモデルを検証するために、核磁気共鳴画像によって脳計測実験を行ったので、その結果に基づきモデルの妥当性を議論する。

キーワード

部分観測, システム同定, オンライン EM アルゴリズム, 埋め込み法, モデル同定強化学習, 脳型情報処理, 核磁気共鳴画像

Partially-observable system identification: applications and a functional brain model*

Wako Yoshida

Abstract

Natural environments sorrounding humans include unobservable states and dynamically change with time. Even in such complicated environments, humans learn the characteristics of the current environment and determine their optimal behaviors. Namely, in a partially-observable environment, humans seem to identify the current environment and solve the optimal decision making problem based on the identification. In this thesis, I discuss the system identification problem in a partially-observable environment on the bases of both machine learning and brain learning.

First, I discuss system identification of nonlinear dynamical systems based on a statistic method. This learning method uses a Normalized Gaussian network and is based on an on-line EM algorithm. In experiments, the NGnet is trained to learn the dynamics of low dimensional chaotic systems, using the time-series of one observable variable. In order to identify the original dynamical system in such a partial observation situation, I employ two kind of embedding method. By using the delay coordinate embedding, which has commonly been used, the NGnet is able to well identify the system dynamics in the delay coordinate space. I propose a new embedding method using infinite impulse response filters, which is called the integral embedding. By investigating the robustness of noise, it is

*Doctor's Thesis, Department of Information Systems, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DT0061025, March 7, 2003.

shown that this integral embedding method is more robust to noise than the previous delay embedding method.

Second, I discuss system identification in a partially-observable environment and an optimal decision making problem using the identified model. In a machine learning field, the optimal decision problem in a stochastic environment is often termed Markov decision process (MDP). If the Markov environment involves unobservable (hidden) state variables, it is formulated as a partially-observable MDP. In this thesis, I propose a model-based reinforcement learning (RL) in an environment with hidden variables. Model-based RL tries to identify the current environment through the past experiences, and makes decisions using the identified model. In my method, the environmental model is estimated based on a Bayes inference with forgetting effect. In order to deal with a dynamic environment that changes with time, I introduce an action selection scheme based on the detection of the environmental change. When applied to a maze task with hidden variables, this method successfully obtains better controls by adapting to the environmental changes better than the existing method.

Third, I propose a possible information processing model in the brain, which realizes the model-based RL above. Recent studies have suggested that the RL algorithm can be associated with the processing of neural systems in the brain. I propose a brain model for RL, in which the major parts of the RL scheme are involved in functions of the prefrontal cortex. I assume the dorsolateral prefrontal cortex executes the maintenance and manipulation of the reward-based environmental model and the anterior prefrontal cortex is related to the estimation of unobsevable states. I also assume the action selection depending on the current environmental model is done within the anterior cingulate cortex. In order to examine this functional model, a human imaging study using functional magnetic resonance imaging is conducted in this study.

Keywords:

partial observation, system identification, on-line EM algorithm, embedding method,

model-based reinforcement learning, brain information processing, functional magnetic resonance imaging

目 次

図 目 次

表 目 次

Chapter 1

序論

1.1. 研究背景

ヒトを取り巻く自然環境は、直接観測できない事象が多く存在する上、時間と共に動的に変化する。このような複雑な環境においても、ヒトは環境の特性を学習し、その変動を予測することによって適切な行動を取ることができる。例えば、車を運転しながら交差点を右折しようとする際には、対向車の動きを予測している。すなわち、近付いてくる対向車の一瞬の観察からその動きを推測し、自車が安全に右折できるかを評価する。野球で外野フライを捕球する際には、一瞬の打球の軌道の観測に基づき落下地点を予測し、後は打球の行方を見ないで走る。このような予測に基づく情報処理は、高等生物、例えばヒトにおいて発達した高度な情報処理と考えられる。この情報処理能力は、ヒトの最も重要な高次機能の 1 つである言語コミュニケーションにおいても重要である。コミュニケーションは、相手との相互作用によって成り立つものであり、そこには何らかの目的が存在する。例えば、相手の言動を無視した一方的な会話や、目的のない言語のやり取りは、コミュニケーションとは言えない。我々は、相手の言動や表情といったコミュニケーション信号に基づいて、相手の心的状態や意図を予測することにより、スムーズなコミュニケーションを行っている。つまり、コミュニケーション能力の高い人は、相手の心的状態を正確に予測し、自らの言動によってそれを自分の目的とする状態へと変化させられる人だといえる。さらに、ヒトの心的状態は時々

刻々と変化するため、その予測も動的に適応できなくてはならない。相手からのフィードバック信号が制約されている電話やＥメールにおいて、しばしばコミュニケーションの不具合いを感じるのは、オンライン的な相手の予測ができないためであると考えられる。

予測を行うためには、観測に基づく対象システムのモデル化が必要である。観測されたデータを用いて未知のシステムをモデル化し、それを用いた予測を行うことは、工学のみならず科学の基礎的な問題であり、例えば、天気予報やシステム制御といった幅広い応用がある。一方で、我々の脳内に構築された制御対象や環境のモデルは、しばしば「内部モデル」と呼ばれている。また、自然環境や工学分野の実問題の多くは、直接観測できない隠れた状態を持つ。このような部分観測状況では、観測可能な情報から隠れた状態を推定し、対象の変化を予測する。例えば、コミュニケーションの相手を対象とした場合、相手の心的状態は直接観測できない隠れ状態であり、言語や表情といった観測可能な時系列から推定する必要がある。脳は、対象システムの内部モデルをシュミレートすることによって、対象の状態や変化を予測し、それに基づいて適切な意思決定を行っている。

以上のように、観測に基づくシステムの同定とその予測問題は、認知科学的な問題としてだけではなく、工学分野での最適制御問題としても興味深いトピックである。

1.2. 本論文の概要

本論文は、対象システムの同定と予測に基づく知的システムについて、システム脳科学の立場で論じようとするものである。ここでシステム脳科学とは、脳の情報処理機構を、アルゴリズム開発を始めとする理論的研究と心理・行動実験によるその検証という実験的研究の両側面から明らかにしようとするものである。

はじめに、非線形力学系のシステム同定問題について、工学的な立場からの手法について議論する。私がここで想定している非線形システムとは、コミュニケーション信号の発信対象である。言語・非言語を問わず、コミュニケーション信号は、強い局所性 (構文構造) と弱い大域性 (意味・文脈構造) を併せ持つという特徴がある。また、このような構造を持つ反面、その時の状況や気分によって異なる

という不安定な時系列であることも重要な特徴である。階層構造は、数理的にはフラクタル性によって特徴付けることができる。本研究では、このコミュニケーション信号の特性が、決定論的な方程式から生成されながら不安定性を含む運動であるカオス軌道の特性と類時していると考え、対象システムとしてカオス力学系を仮定する。すなわち、脳は、フラクタル性と不安定性を持つ時系列であるコミュニケーション信号に基づいて対象システムである相手の心的状態を同定し、それを用いた予測を行うことで、スムーズなコミュニケーションを実現していると仮定する。また、対象システムの次数 (内部変数の数) よりも、観測信号の次数は少ない場合がしばしばある。直接観測できない変数のことを隠れ変数と呼び、隠れ変数がある状況で対象システムの同定を行う必要がある。以上のことから、コミュニケーションは部分観測環境におけるシステム同定とそれに基づく最適意思決定問題とみなすことができる。本論文の第 2章では、隠れ変数を持つ未知のカオス力学系のシステム同定を行い、それを用いて時系列の将来を予測する工学的手法について論じる。この研究は工学的研究であり、シミュレーション実験によって従来の提案手法よりも大幅に優れた結果が示されるが、上述のように、その動機付けはシステム脳科学によるものである。

次に、環境の同定とそれに基づく予測を用いた最適意思決定のアルゴリズムに関して論じる。最初に述べた車の右折の例にもあるように、我々は、観測に基づき環境あるいは対象を同定し、同定された内部モデルに基づいて変化を予測し、その予測を用いることで意思決定を行っている。意思決定における最適性が、環境からのフィードバック信号、特にスカラー信号によって定義されていることを仮定する。このスカラー信号のことを報酬 (あるいは罰) と呼ぶ。すなわち、報酬によって最適性が決まっていることを仮定する。さらに、環境が無記憶の確率過程でモデル化できることを仮定した場合に、こうした報酬依存の最適化問題は、しばしばマルコフ決定過程と呼ばれている。ここで、報酬依存の意思決定は、動物において最もプリミティブかつ主要な学習法である点に注意する。心理学の実験でよく用いられるスキナー箱は、動物が自発的に中にあるレバーを押すとエサが与えられる仕組みになっているものである。箱に入れられた空腹のネズミは、試行錯誤の後に、レバーを押す行動を獲得する。心理学では、こうした行動と報酬の対連合学習をオペラント条件付けと呼ぶ。オペラント条件付けタスクは報酬

依存の最適化問題である。一方、工学的には、報酬依存の最適化問題、すなわちマルコフ決定過程の解法として、強化学習が提案されている。本論文の第 3章では、未知の環境を観測に基づき同定しながら、最適な意思決定を行う工学的手法について、強化学習の定式化に基づき議論する。特に、環境の変化に基づき、行動の貪欲さとランダムさを適切に制御する手法について詳しく議論する。この研究もまた工学的な研究であるが、上述のように、脳における意思決定法に動機付けがある。

そこで、本論文の第 4章において、こうした工学的な強化学習法が脳においていかに実現されているかに関して私の仮説を述べる。元来強化学習は、心理学実験で観察される動物の行動変容からヒントを得たものである。また、神経生理学研究によって、脳内で強化学習法と類似の機構が働いていることが示唆されている。ネズミがレバーを押したときに、脳のある部位に電気信号を送るようにすると行動の強化が起こり、ネズミはレバーを押し続けるようになる。このような現象を自己刺激と呼び、視床下部を中心とした辺縁系で起こる。このうち、ノルアドレナリン作動性ニューロンは得られる報酬自体に反応して発火するのに対し、ドーパミン作動性ニューロンは報酬を予測する事象に反応して発火する [69]。この活動は、それぞれ強化学習における直接的報酬および報酬予測の誤差の振舞いと類似している。また、サルの運動前野に、他のサルや人間がある行為をしている場面を観察したときと、自分が同様の行為をしているときの両方において活動する細胞、すなわちミラーニューロンが存在することが明らかになった [57]。このミラーニューロンは、他者の行為を自分自身の行為と対応付ける役割を持つ可能性があり、これは脳内に対象の内部モデルが存在することを示唆する。これらの知見は、モデル同定強化学習法が脳内で実現されている可能性を示唆しており、第 4章ではその脳型情報処理モデルを提案する。第 3章で述べた手法では、未知の環境を同定し、それを用いて報酬の予測を行い、行動を選択する必要がある。また、環境の変化を認識し、行動選択の戦略を切替える必要がある。さらに、環境に存在する隠れ変数を観測に基づき推定する必要がある。こうした機能が、脳のいかなる領域によって実現されているかを、現在までの神経生理学、認知心理学の知見に基づき、論じる。

最後に、第 4章で述べた仮説を検証するための認知心理学的研究を行ったの

で、本論文の第 5章において論じる。ヒトに強化学習タスクを行わせ、その間の脳活動を非侵襲脳活動計測装置、特に機能的核磁気共鳴図 (functional magnetic resonance imaging: fMRI) を用いて計測を行った。その結果、強化学習を遂行するのに必要な脳の機能部位の特定ができ、第 4章で述べた仮説の一部を検証することができた。

このように工学的手法の研究からの知見を実際に脳のモデルとして構成し、それを認知心理学的に検証していくこと、これがシステム脳科学の手法である。本論文の第 6章では、システム脳科学としての私の一連の研究のまとめとして、現在行っている研究について論じる。1 つは、環境の同定を行いながら強化学習を実行する工学的手法に関する最新の研究について述べる。第 3章で述べた手法は、環境が離散的であることを仮定していたが、最新の研究では連続性を仮定している。実世界は連続的であるので、新しい手法の方が現実的であり、本論文の第 2章で述べた連続システムの同定法との関連もより強いものとなる。第二に、環境に隠れた変数がある場合の強化学習を遂行する脳の機能局在について、将来の研究構想を含めて論じる。

1.3. 論文の構成

本論文の構成を紹介する。

第 2章では、非線形力学系の部分観測システム同定に関する研究について述べる。未知の非線形力学系の状態変数のうち 1 変数の時系列のみが観測される状況において、教師あり学習によってシステムを同定する。学習器としては、正規化ガウス関数ネットワークという一種のニューラルネットワークを用い、統計的学習法によって関数近似を行う。はじめに、2.1節で本研究で用いた学習器とその学習法について説明する。2.2節では、学習対象として用いたカオス力学系とそのシステム同定手法について概説する。1 次元の観測時系列から元の力学系を再構成する手法として、埋め込み法を用いる。2.3節では、従来用いられてきた遅れ座標埋め込み法と、本研究で提案する積分埋め込み法について説明する。2.4節で、本手法を用いてシミュレーション実験を行った結果を報告し、考察を行う。

第 3章では、部分観測環境におけるモデル同定強化学習について議論する。3.1

節で強化学習について概説し、3.2節と 3.3節ではそれぞれ、本研究で用いた環境モデルの同定手法と行動選択の制御手法について説明する。環境モデルの近似にはベイズ推定を用いる。また、エージェントの行動選択に逆温度メタパラメータの制御機構を導入する。この制御は、行動価値関数のばらつきと環境変化の認識の両方を基づいて行う。3.4節で本研究で用いたアルゴリズムをまとめ、3.5節で本手法を迷路探索問題に適用した結果を報告する。

第 4章では、モデル同定強化学習の脳型情報処理モデルを提案し、第 5章でその検証実験について報告する。5.1節で非侵襲脳計測機器による検証実験の概要を説明する。5.2節で実験データの解析結果を報告し、考察する。

第 6章で、これらの研究をまとめ、さらに今後の課題について議論する。

Chapter 2

非線形力学系の部分観測システム同定

ヒトは複雑な時系列から対象システムの同定を行い、それを用いた環境変化の予測を基に意思決定を行っていると考えられる。こうした高度な情報処理は、特にコミュニケーションにおいて重要になる。コミュニケーション信号は、長距離依存構造を持つという規則的な側面と、状況や気分によって異なるという不確定な側面を持ち、これらの特性はカオス軌道の特性と類似していると考えられる。近年、生体から得られる多くの信号がカオス的な挙動を示すことが明らかとなり、ヒトの音声や言語のカオス性を論じた研究も多数行われている。ヒトの音声波形にはピッチ間隔や振幅や波形に関するゆらぎがあり、どの一周期をとってみても全く同じものはない。そして、このゆらぎがヒトの音声が自然に聞こえるための重要な要素になっていることが指摘されている。発話音声から生成されるストレンジアトラクタの力学的性質を調べた研究は、平常時と疲労時とでリアプノフ指数の値に大きな差が現れることを示した。また、自然言語における単語の出現頻度や音楽の周波数が 1/f の関係で分布することが知られている。これらは、言語や音楽といったコミュニケーション信号がカオス的なゆらぎによって特徴付けられることを示唆するものである。一方で、コミュニケーション信号である言語をカオスを用いてモデル化しようとする研究も多数行われている。カオスニューラルネットワークを用いた研究は、ネットワークに様々な文字を学習させると、一種の言語のように覚えた文字から成る時系列を生成することを示した。以上のことから、本章では、コミュニケーション信号をカオス時系列であると仮定し、学習対象としてカオス力学系を用いる。

また、実問題の多くは直接観測できない隠れ変数を持つ部分観測状況であり、コミュニケーションにおいても、相手の心的状態は隠れ変数とみなすことができる。本章では、カオス力学系の 1 つの力学変数の軌道の観測に基づき、システム同定を行う統計的学習アルゴリズムを開発し、シミュレーションにより評価を行う。

2.1. モデルと学習法

本研究では、学習データとして未知の対象システムの入出力データ集合が与えられる、教師あり学習を扱う。教師あり学習によってシステム同定を行うための関数近似モデルは、大域モデルと局所モデルに分類することができる。多層パーセプトロンに代表される大域モデルは、観測データ空間全体を全てのモデルパラメータを用いて近似するため、学習速度が遅くなる。一方、観測データ空間を領域分割し各部分空間で近似を行うモデルを局所モデルと呼び、動径基底関数 (Radial Basis Function: RBF) や Mixture of Experts network がある。局所モデルは、1 つのデータを学習するためには少数のモデルのパラメータを変えるだけでよく、学習が容易である点が長所である。しかし、局所モデルでデータ空間の全域にわたって近似を行おうとすると、必要な部分空間の数が入力空間の次元の増加に伴って指数的に増大し、その結果として計算量の爆発が起こるという短所がある。どちらのモデルを適用するのが良いかは、対象とする問題に依存する。しかし、実際のデータは力学系のアトラクタのように入力空間よりも低い次元に分布することが多く、学習に必要なリソースをそうした部分空間に主に配備すれば、計算量の爆発は起こりにくい。本研究で考える問題はこのような場合に相当するため、局所モデルが適当であると考えられる。

ヒトの学習法を考えた場合でも、複雑な環境モデル化する場合に 1 つ 1 つの入力状態に対してモデル全体を更新するとは考え難く、複数の部分モデルの集合によって表現しているというが尤もらしい。環境モデルを部分化しておくことにより、類似した複数の環境モデルを効率良く保持することができるという利点もある。また、ヒトを取り巻く環境は時間と共に動的に変化するため、実時間のモデル同定が必要となり、高速なオンライン学習が必須である。

そこで、本研究では、関数近似器として局所モデルの一種である正規化ガウス関数ネットワーク (Normalized Gaussian network: NGnet) [47] を用い、オンライン EM アルゴリズムによって学習を行う。NGnet [47] は局所的に線形近似を行うユニットからなるネットワークである。このモデルは正規化ガウス関数を用いて入力空間を柔らかく領域分割し、それぞれのユニットが各領域において出力を線形近似する。

2.1.1 正規化ガウス関数ネットワーク

N 次元入力ベクトル x を D 次元出力ベクトル y に変換する NGnet は以下の方程式で定義される。

$$y = \sum_{i=1}^{M}\left(\frac{G_i(x)}{\sum_{j=1}^{M} G_j(x)}\right)(W_i x + b_i) \tag{2.1a}$$

$$G_i(x) \equiv (2\pi)^{-N/2}|\Sigma_i|^{-1/2} \times \exp\left[-\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)\right] \tag{2.1b}$$

M はユニットの個数、プライム記号 ($'$) は転置を表す。$G_i(x)$ は N 次元中心位置ベクトル μ_i と $N \times N$ 次共分散行列 Σ_i をパラメータとして持つガウス関数である。$|\Sigma_i|$ は Σ_i の行列式である。W_i と b_i はそれぞれ $D \times N$ 次線形回帰行列と D 次元バイアスベクトルである。以下では $\tilde{W}_i \equiv (W_i, b_i), \tilde{x}' \equiv (x', 1)$ の表記法を用いる。

図 2.1に、1 次元ガウス関数を用いた NGnet の関数近似過程を示す。図の横軸と縦軸はそれぞれ、1 次元入力と 1 次元出力を表す。NGnet は、入力空間を $M(=4)$ 個の正規化ガウス関数ユニットで領域分割する (図 2.1(a))。学習データとして入出力データが与えられると、各ユニットがそれぞれの領域内で線形近似を行う (図 2.1(b))。この線形行列と正規化ガウス関数の積が各ユニットの出力となり (図 2.1(c))、その総和がネットワークの出力となる (図 2.1(d))。NGnet の学習とは、入出力データをうまく近似できるように、各ユニットの中心位置や分散を更新することである。

NGnet は入力変数 x と出力変数 y の対 (x, y) を確率的 (不完全) 事象とする確率モデルとみなすことができる [32]。各観測事象 (x, y) に対しユニットの集合 $\{i|i = 1, ..., M\}$ から 1 つのユニットが選ばれるものと仮定する。ユニット番号 i

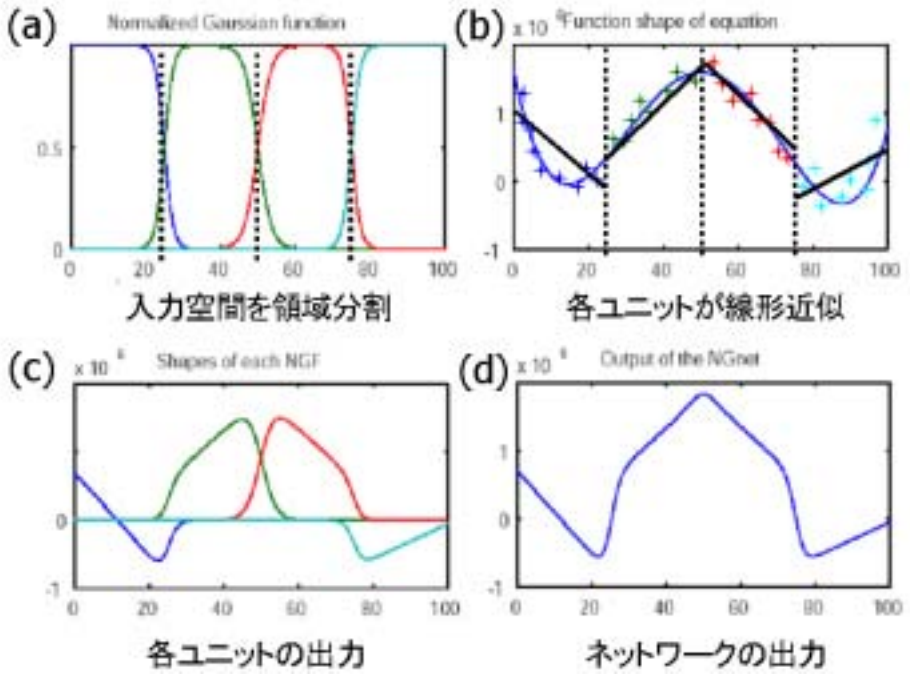

Figure 2.1 正規化ガウス関数ネットワークの関数近似過程

は隠れ変数とみなされる。(x, y, i) の 3 つ組を完全事象と呼ぶ。この確率モデルは 1 つの完全事象に対する以下の確率分布によって定義される [91]。

$$P(x, y, i|\theta) = \frac{G_i(x)}{M}(2\pi)^{-\frac{D}{2}}\sigma_i^{-D}\exp\left[\frac{-1}{2\sigma_i^2}(y - \tilde{W}_i\tilde{x})^2\right] \tag{2.2}$$

ここで $\theta \equiv \{\mu_i, \Sigma_i, \sigma_i^2, \tilde{W}_i | i = 1, ..., M\}$ はモデルパラメータの集合である。確率分布 (2.2) より、与えられた入力変数 x に対する出力変数 y の期待値が求められる。

$$E[y|x] \equiv \int yP(y|x, \theta)dy = \sum_{i=1}^{M}\left(\frac{G_i(x)}{\sum_{j=1}^{M} G_j(x)}\right)(\tilde{W}_i\tilde{x}) \tag{2.3}$$

これは NGnet の出力 (2.1) と等価である。すなわち確率分布 (2.2) は NGnet の確率モデルを与える。

2.1.2 オンライン EM アルゴリズム

T 個の観測事象の組 $(X,Y) \equiv \{(x(t),y(t))|t=1,...,T\}$ から確率モデル (2.2) のパラメータ θ を最尤推定法により決めることができる。特に EM アルゴリズム [20] は隠れ変数を持つモデルに適用することができる。EM アルゴリズムは以下に述べる E ステップと M ステップを繰り返すものである。E ステップと M ステップを 1 回ずつ行うと観測データの組に対する尤度が増大する (または変化しない) ので、E ステップと M ステップを繰り返すことによって最尤推定量が漸近的に求まる。

- E (Expectation) ステップ

  現在の推定値を $\bar{\theta}$ とする。この $\bar{\theta}$ を用いて各観測データ $(x(t),y(t))$ に対して i 番目のユニットが選ばれる事後確率をベイズ則によって計算する。

$$P(i|x(t),y(t),\bar{\theta}) = \frac{P(x(t),y(t),i|\bar{\theta})}{\sum_{j=1}^{M} P(x(t),y(t),j|\bar{\theta})} \tag{2.4}$$

- M (Maximization) ステップ

  事後確率 (2.4) を用いた完全事象に対する期待対数尤度 $Q(\theta|\bar{\theta},X,Y)$ を以下のように定義する。

$$Q(\theta|\bar{\theta},X,Y) = \sum_{t=1}^{T}\sum_{i=1}^{M} P(i|x(t),y(t),\bar{\theta}) \log P(x(t),y(t),i|\theta) \tag{2.5}$$

  $Q(\theta|\bar{\theta},X,Y)$ の増大は観測データの組に対する尤度の増大を意味する [20] ので、$Q(\theta|\bar{\theta},X,Y)$ を推定値 θ に関して最大化する。そこで条件式 $\partial Q/\partial\theta = 0$ を解くことにより、新しいパラメータの推定値が以下のように求められる [91]。

$$\mu_i = \langle x\rangle_i(T)/\langle 1\rangle_i(T) \tag{2.6a}$$

$$\Sigma_i^{-1} = [\langle xx'\rangle_i(T)/\langle 1\rangle_i(T) - \mu_i(T)\mu_i'(T)]^{-1} \tag{2.6b}$$

$$\tilde{W}_i = \langle y\tilde{x}'\rangle_i(T)[\langle \tilde{x}\tilde{x}'\rangle_i(T)]^{-1} \tag{2.6c}$$

$$\sigma_i^2 = \frac{1}{D}\Big[\langle|y^2|\rangle_i(T) - Tr\Big(\tilde{W}_i\langle\tilde{x}y'\rangle_i(T)\Big)\Big]/\langle 1\rangle_i(T) \tag{2.6d}$$

ここで $Tr(\cdot)$ は行列の対角和を表す。また $\langle\cdot\rangle_i$ は事後確率 (2.4) に関する重み付き平均値であり、以下で定義される。

$$\langle f(x,y)\rangle_i(T) \equiv \frac{1}{T}\sum_{t=1}^{T} f(x(t),y(t))P(i|x(t),y(t),\theta) \tag{2.7}$$

(2.6a) 式は、ガウス関数の中心位置が、事後確率に基づき各ユニットの領域に属するとされた入力データの重み付き平均値で与えられることを意味する。同様に (2.6b) 式は、ガウス関数の共分散行列が、そのユニットの領域に属するとされた入力データの重み付き共分散で与えられることを意味する。(2.6c) 式は、線形係数が属するデータから重み付き最小二乗法で求められることを意味する。

上で述べた EM アルゴリズムは、パラメータを更新する度に全観測データ (X,Y) を用いるバッチ学習 [91] に基づくものである。ここでは EM アルゴリズムのオンライン学習法 [65] を示す。この学習法では各観測データが与えられるたびに推定値を変更するので、t 番目のデータ $(x(t),y(t))$ が与えられた後の推定値を $\theta(t)$ とすることにする。また忘却係数 $\lambda(t) \in [0,1]$ を導入し、重み付き平均値 $\ll f(x,y) \gg_i (t)$ を以下のように定義する。

$$\begin{aligned} \ll f(x,y) \gg_i (t) &\equiv \eta(t)\Bigg[\sum_{\tau=1}^{t}\left(\prod_{s=\tau+1}^{t}\lambda(s)\right) f(x(\tau),y(\tau)) \\ &\qquad \times P(i|x(\tau),y(\tau),\theta(\tau-1))\Bigg] \qquad (2.8a) \\ \eta(t) &\equiv \left(\sum_{\tau=1}^{t}\left(\prod_{s=\tau+1}^{t}\lambda(s)\right)\right)^{-1} \qquad (2.8b) \end{aligned}$$

オンライン EM アルゴリズムは、この重み付き平均値を逐次的に計算することにより得られる。まず、E ステップでは現在のデータ $(x(t),y(t))$ に対して事後確率 $P_i(t) \equiv P(i|x(t),y(t),\theta(t-1))$ を (2.4) 式を用いて計算する。M ステップでは重み付き平均値 $\ll 1 \gg_i (t), \ll x \gg_i (t), \ll |y|^2 \gg_i (t), \ll y\tilde{x}' \gg_i (t)$ を、(8) 式の定義を展開した以下の逐次計算式を用いて計算する。

$$\begin{aligned} \ll f(x,y) \gg_i (t) &= \ll f(x,y) \gg_i (t-1) + \\ &\qquad \eta(t)[f(x(t),y(t))P_i(t) - \ll f(x,y) \gg_i (t-1)] \qquad (2.9a) \\ \eta(t) &= (1+\lambda(t)/\eta(t-1))^{-1} \qquad (2.9b) \end{aligned}$$

次に、新しい推定値 $\theta(t)$ を以下のように求める。

$$\mu_i(t) = \ll x \gg_i (t) / \ll 1 \gg_i (t) \tag{2.10a}$$

$$\tilde{\Lambda}_i(t) = \frac{1}{1-\eta(t)}\left[\tilde{\Lambda}_i(t-1) - \frac{P_i(t)\tilde{\lambda}_i(t-1)\tilde{x}(t)\tilde{x}'(t)\tilde{\Lambda}_i(t-1)}{(1/\eta(t)-1) + P_i(t)\tilde{x}'(t)\tilde{\Lambda}_i(t-1)\tilde{x}(t)}\right] \tag{2.10b}$$

$$\tilde{W}_i(t) = \tilde{W}_i(t-1) + \eta(t)P_i(t)(y(t) - \tilde{W}_i(t-1)\tilde{x}(t))\tilde{x}'(t)\tilde{\Lambda}_i(t) \tag{2.10c}$$

$$\sigma_i^2(t) = \frac{\ll |y|^2 \gg_i (t) - Tr\left(\tilde{W}_i(t) \ll \tilde{x}y' \gg_i (t)\right)}{D \ll 1 \gg_i (t)} \tag{2.10d}$$

ここで $\tilde{\Lambda}_i(t) \equiv [\ll \tilde{x}\tilde{x}' \gg]_i^{-1}(t)$ である。$\Sigma_i^{-1}(t)$ は以下の関係式を用いて $\tilde{\Lambda}_i(t)$ から求められる。

$$\tilde{\Lambda}_i(t) \ll 1 \gg_i (t) = \begin{pmatrix} \Sigma_i^{-1}(t) & -\Sigma_i^{-1}(t)\mu_i(t) \\ -\mu_i'(t)\Sigma_i^{-1}(t) & 1 + \mu_i'(t)\Sigma_i^{-1}(t)\mu_i(t) \end{pmatrix} \tag{2.11}$$

このアルゴリズムでは、補助変数 $\tilde{\Lambda}_i(t)$ の導入により共分散行列 $\Sigma_i(t)$ の逆行列を計算する必要がない。

実質的な学習係数である $\eta(t)$ について以下の条件を満たすようなスケジューリングを行うと、オンライン EM アルゴリズムが最尤推定量を求めるための確率近似法を実現していることを示すことができる [65]。

$$\eta(t) \stackrel{t\to\infty}{\longrightarrow} 0, \quad \sum_{t=1}^{\infty}\eta(t) = \infty, \quad \sum_{t=1}^{\infty}\eta^2(t) < \infty \tag{2.12}$$

例えば、忘却係数 $\lambda(t)$ について、$1-\lambda(t) \stackrel{t\to\infty}{\longrightarrow} o(1/t)$ のように 1 に収束するスケジューリングを行えば、(2.12) 式の条件が満足され、オンライン EM アルゴリズムは必ず局所最尤解に収束することが分かる。

2.1.3 ユニットの動的操作

現実のデータは力学系のアトラクタのように入力空間の次元よりも低い次元の入力分布を持つことが多い。こうした場合に、入力データが出現するところにのみユニットを配置すれば十分であるというのが局所モデルの長所の一つである。できるだけデータの入出力分布にそったユニット配置を行うためには、データの

出現に応じてオンライン的にユニットを生成させたり、消滅させたりする機構があることが望ましい。本研究では、データが発生する領域に効率良くユニットを配置するために、オンライン的にユニットを生成消滅させる機構を導入する。これらの機構はデータの入出力分布が時間と共に変化しているような状況を扱う上でも有効である [65]。

- ユニットの生成

  新しいデータ $(x(t), y(t))$ が与えられた時、現在のパラメータ値 $\theta(t-1)$ を用いたモデルがこのデータを発生させる確率は $P(x(t), y(t)|\theta(t-1)) \equiv \sum_{i=1}^{M} P(x(t), y(t), i|\theta(t-1))$ で与えられる。$0 < P_{\text{produce}} \ll 1$ とすると、$P(x(t), y(t)|\theta(t-1)) < P_{\text{produce}}$ のとき、このデータは現在の全てのユニットから離れすぎているためにモデルによって説明される確率が小さいことになる。この場合、このデータを担当する新しいユニットが生成される。

- ユニットの削除

  (2.9) 式によって求められる重み付き平均値 $\ll 1 \gg_i (t)$ は i 番目のユニットが時刻 t までに与えられたデータを説明するのにどの程度使われたかを示す。$0 < P_{\text{delete}} \ll 1/M$ とすると、$\ll 1 \gg_i (t) < P_{\text{delete}}$ はこのユニットが殆んど使われていないことを表す。この場合、ユニット i を削除する。

- ユニットの分裂

  ユニットの誤差分散 $\sigma_i^2(t)$(2.10d 式) はユニット i の予測誤差の期待値を表す。D_{divide} をある正値とする。$\sigma_i^2(t) > D_{\text{divide}}$ であるとき、線形近似を担当している範囲が広すぎるために、このユニットの予測が不十分であると考えられる。この場合、ユニット i を 2 つに分割し担当する領域も半分にする。

なお、学習の第 1 ステップでは時系列の最初の 1,000 点について k 平均法を用いてクラスタリングを行い、各クラスタ内において等方分散ガウス分布の推定を行った。線形回帰行列については、各クラスタ内においてバイアスベクトルを最小二乗法を用いて求め、線形係数は 0 とした。この初期化は、最初の 1,000 点について、NGnet のパラメータに対して $\Sigma_i = \sigma_{x,i}^2 \cdot I_N, \sigma_i^2 = \infty, W_i = 0$ の制限条件

を付けてオンライン EM アルゴリズムを実行することに対応する。ただし I_N は N 次元単位行列である。学習が進行すると、データ分布に適応して、ユニットの分散行列はしばしば異方なものとなる。

2.2. カオス力学系の学習

2.2.1 カオス力学系

本研究では、NGnet にカオス力学系のダイナミクスを学習させることを試みた。力学系に初期条件を与え、十分時間が経った後の漸近的な振舞いを力学系のアトラクタという。力学系のアトラクタは、平衡点、リミットサイクル、トーラス、ストレンジアトラクタに分類され、このストレンジアトラクタがカオス力学系を特徴づけるアトラクタであり、しばしばカオスアトラクタと呼ばれる。カオス軌道はアトラクタ上で不安定であるため、長時間にわたる時間的振る舞いを学習させることは不可能である。しかし、カオス軌道は常に同じ方程式によりその時間発展が決められるので、カオスダイナミクスを巨視的に表現しているカオスアトラクタを学習させることは可能であると考えられる。

実験には低次元カオスとして代表的な 2 種類のカオスシステムを用いた。1 つは以下の微分方程式によって定義されるレスラー (Rössler) システムである。

$$\begin{pmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{pmatrix} = \begin{pmatrix} -y - z \\ x + \alpha y \\ \beta x - \gamma z + xz \end{pmatrix} \tag{2.13}$$

ここで $\dot{x} \equiv dt/dx$ は変数 x の時間微分を表し、レスラーシステムの各パラメータは $\alpha = 0.36, \beta = 0.4, \gamma = 4.5$ とした [44]。また、以下の微分方程式によって定義されるローレンツ (Lorenz) システムも用いた。

$$\begin{pmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{pmatrix} = \begin{pmatrix} \alpha(y - x) \\ -y + (\beta - z)x \\ -\gamma z + xy \end{pmatrix} \tag{2.14}$$

ここで、ローレンツシステムの各パラメータは $\alpha = 10.0, \beta = 28.0, \gamma = 8/3$ とした。レスラー方程式 (2.13) の軌道は時間間隔 $\delta t_R = 0.01$、ローレンツ方程式 (2.14)

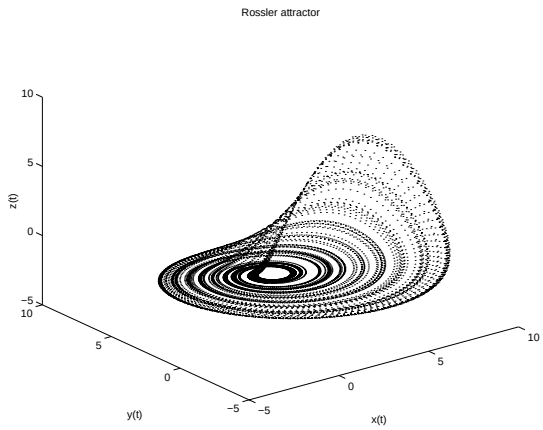

Figure 2.2 レスラーアトラクタの相図

の軌道は時間間隔 $\delta t_R = 0.001$ の 4 次の Runge-Kutta 法を用いて求めた。学習データはこれらの軌道から、レスラーシステムでは時間間隔 $\delta t_O = 0.05$、ローレンツシステムでは $\delta t_O = 0.01$ でサンプルされるものとする。図 2.2および図 2.3 に各時間間隔でサンプルされたアトラクタの相図を示す。

カオス軌道はあるプロセスを不規則に繰り返しながら、非周期的で複雑な挙動を示す。例えばレスラー軌道は以下のようなプロセスを繰り返す。

1. $z = 0$ 平面近傍に拘束されながら原点の回りを数回転する。

2. z 方向に立ち上がりながら原点の方向へ折り返す

3. 原点近くに再び引き戻された後、$z = 0$ 平面近傍に拘束される。

この奇妙な軌道は、初期値を変えても現れるためアトラクタであるということが分かる。これがカオスアトラクタである。このようなカオスアトラクタは初期値に鋭敏に反応する。初期値に対する鋭敏な依存性はバタフライ効果とも呼ばれ、微小な誤差の影響が時間と共に指数関数的に拡大される性質を言う。初期値に限らず、微小な外乱によっても同様の性質が生じ、それは軌道不安定性とも呼ばれる。実世界では無限の精度での観測や初期値の指定などは不可能であるので、システムがカオスであるなら、長期の予測は原理的に不可能となる。

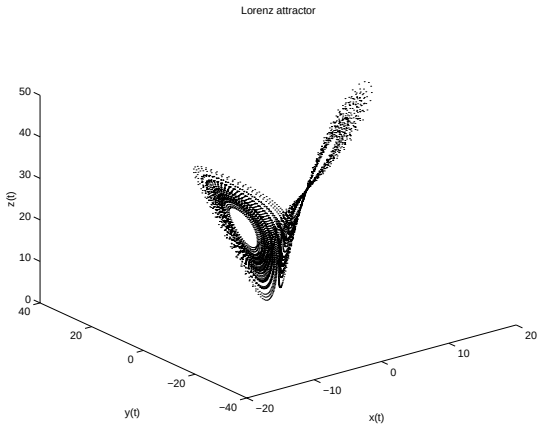

Figure 2.3 ローレンツアトラクタの相図

一方でカオスは、微小な外乱によって軌道不安定性が生じても、状態空間において定常的振舞いを表すアトラクタの幾何学的構造は変化しないという安定性を有する。カオス軌道は常に同じ方程式によりその時間発展が決められるという決定論的な法則に従うため、カオスダイナミクスを巨視的に表現しているカオスアトラクタを近似させることは可能であると考えられる。本研究の目的は、NGnetにこのカオスアトラクタを学習させることであるが、実際には 3 次元のアトラクタそのものを学習させるわけではなく、そのダイナミクスを学習させることにある。

2.2.2 カオス力学系の再構成

本研究では、NGnet とオンライン EM アルゴリズムを非線形力学系のシステム同定問題に適用する。関数近似器である NGnet は、時系列を教師信号とし、その時系列を生成するダイナミクスを近似するように学習を行う。図 2.4に学習の概要を示す。$\dot{X}$ を状態変数 X の時間微分としたとき、力学方程式 $\dot{X} = F(X)$ を持つ未知の力学系を想定する。この力学系から観測された観測変数の時系列を用いて NGnet の学習を行う。学習時に与えられるデータは入力とターゲット出力の対であり、NGnet はこの入出力関係を近似するようにトレーニングされる。もし学習後の NGnet がダイナミクスを獲得していれば、初期状態を与えることによっ

て自動的に軌道を生成することができる。この生成軌道と元の力学系からの軌道が類似していれば、NGnet が未知の力学系を再構成することができたと言える。

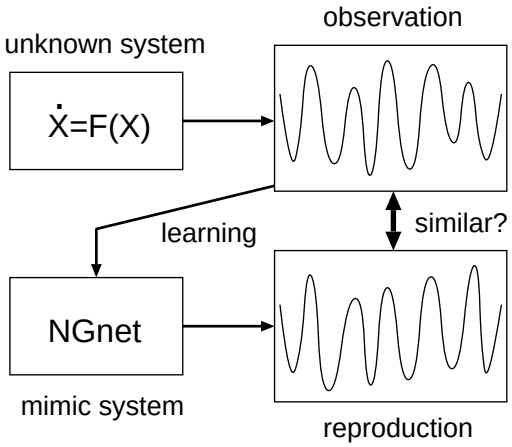

Figure 2.4 時系列データを用いた学習

NGnet は入出力対である学習データから入出力関係を近似するように学習を行うが、本研究では 2 種類の入出力関係を定義する。1 つは力学系の離散化マップを学習する手法で、もう 1 つは力学系の離散化ベクトル場を学習する手法である。離散化マップ学習では、現在の状態ベクトルを入力とし、次の観測時刻における状態ベクトルを出力とする。一方、離散化ベクトル場学習では、現在の状態ベクトルと、現在の状態と次の観測時刻における状態の差分の対を学習データとして与える。

あるシステムのベクトル表記を $\dot{X} = F(X)$ とする。離散化マップ学習では、NGnet は時間間隔 δt_O での離散化マップ、$X(t + \delta t_O) = K(X(t))$ を近似するように学習を行う。つまり、NGnet の入力は現在の状態ベクトル $X(t)$、出力は次のサンプリング時刻での状態ベクトル $X(t + \delta t_O)$ をターゲットとして学習を行う。ここでシステムは、写像 K で表現される。学習後、NGnet が離散化マップ K を獲得していれば、その近似を繰り返すことによって未知のシステムによる軌道を生成することができる。離散化ベクトル場学習は、NGnet の入力としては現在の状態ベクトルを与え、出力としてアトラクタのベクトル場、すなわち状態の微分をターゲットとして学習を行う。システムの微分方程式を時間間隔

δt_O で離散化して、NGnet の入力は状態ベクトル $X(t)$、出力は離散化ベクトル場 $F_O(X(t)) \equiv (X(t+\delta t_O) - X(t))/\delta t_O$ をターゲットとして学習を行う。一方で学習後にシステムを動作させる場合には、現在の状態 $X(t)$ における離散化ベクトル場が NGnet の出力として計算される。次の時刻の状態 $X(t+\delta t_O)$ はこのベクトル場を用いて計算される。この学習法は、離散化マップをターゲットとする学習法と違って微分を用いるため、時間的なハイパスフィルターを掛けていることになる。本研究では、両方の学習法を試した上で、離散化ベクトル学習を用いた場合の結果について報告する。

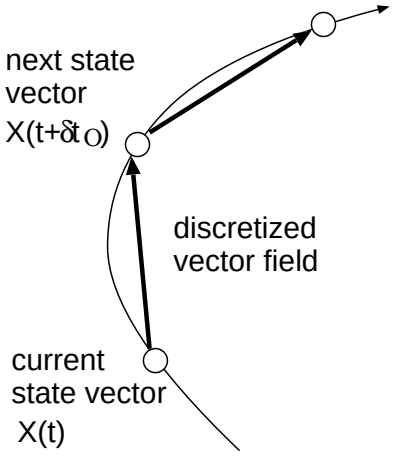

Figure 2.5 離散化ベクトル場の学習

2.2.3 定量的指標

カオスの特徴には、軌道不安定性、長期予測不能性、自己相似性があり、これらの特徴はそれぞれ、リアプノフ指数、KS エントロピー、各種のフラクタル次元で定量的に評価することができる。本研究では、学習後に生成された軌道がカオスであるかを判定し、元のカオスアトラクタをどの程度近似できているかを定量的に評価するために、相関次元 [29] とリアプノフ指数 [23] という 2 つの指標を用いた。これらの指標は、カオスの異なる特徴を力学的および幾何学的に定量化する指標であり、力学系の時間発展の下で不変である。従って軌道の初期状態

に関わらず一定であるため、初期値鋭敏性という特徴を持つカオスを扱う上で有効である。以下に各指標の概要を述べる。

- 相関次元 (correlation dimension) [29]

  相関次元は、カオスアトラクタの幾何学的複雑度を示す量であり、フラクタル次元の一種である。n 次元の観測時系列を $\{X_j | j = 1, 2, \cdots, T\}$ とすると、距離 r に対する相関関数 $C(r)$ は以下のように定義される。

  $$C(r) = \lim_{T \to \infty} \frac{1}{T^2} \sum_{i=1}^{T} \sum_{j=1}^{T} \theta(r - \|X_i - X_j\|) \tag{2.15}$$

  ここで $\theta(x) = 1(x \geq 0); = 0(x < 0)$ で、$\|\cdot\|$ はノルムである。つまり (2.15) 式は定めた距離 r よりも近い所にある n 次元ベクトル X_j の対の数の割合を表す。いくつかの r の値で計算した $C(r)$ を r に対して両対数プロットするとその勾配から ν が求まる。これが相関次元である。もし X_j が n 次元空間内で一様に分布しているなら $\nu = n$ となる。カオスアトラクタの次元は有限であり、実数値である。

- 最大リアプノフ指数 (maximum Lyapunov exponent) [23]

  リアプノフ指数は、近接した 2 点から出発した 2 つの軌道の時刻 $n \to \infty$ での乖離度を測定したものであり、カオスの力学的複雑度を特徴づける量の一つである。一般にカオス力学系では、不安定方向と安定方向が存在する。2 次元力学系に初期値の集合として微小円を与えたものを仮定する。最初は円であったものが、1 回写像されることによって、例えば、縦方向には引き延ばされ、横方向には押し潰される結果、楕円となる。このとき、縦 (横) の各方向に対する指数的拡大 (縮小) 率、λ_1, λ_2 を考えることができ、これらをリアプノフ指数と呼ぶ。リアプノフ指数が負または 0 になる場合は、系が準周期性を示す指標となる。またリアプノフ指数の和は体積の拡大率を表すため、散逸系では負になる。力学系がカオスを生じるかどうかを判定するためには、リアプノフ指数のうち、最大のものが正であるかどうかを調べればよい。

$x(t)$ を 1 次元空間内の時刻 t の座標とし、その写像を $f(x(t))$ とすると、リアプノフ指数 λ は、以下のように定義される。

$$\lambda = \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} \log |f'(x(t))| \tag{2.16}$$

一般的な多次元系においては d 次元の離散力学系を考える。$x(t)$ における微小変位 $\Delta x(t)$ に関する写像はヤコビ行列として求める。このヤコビ行列の N 回分の積行列から正定値行列を定義し、その固有値からリアプノフ指数を求めることができる。力学系が実際に分かっている場合、リアプノフ指数の推定は比較的簡単であるが、それでも、正定値行列の定義通りに計算することは容易ではない。これは、初期変位として与えた $\Delta x(0)$ がカオスアトラクタの安定多様体方向に押し潰されるため、数値計算上は安定方向の負のリアプノフ指数の評価が困難となるためである。そこで、実際の数値計算では、N 回の各反復毎に $\Delta x(t)$ が伸びた (縮んだ) 結果を正規直交化することにより、この問題を避ける。直交化の手法としては、QR 分解や特異値分解があるが、本研究では、グラム-シュミットの直交化法を用いた。

2.2.4 システムノイズと観測ノイズ

現実のデータにはノイズが含まれる。本研究で用いるシステムは人工システムであるが、ノイズを含む現実のデータの学習におけるノイズの影響を調べるために、人工的に 2 種類のノイズを付加し、それらの影響について検討した。ここで 2 種類とはシステムのダイナミクスそのものを乱すシステムノイズと、システムの力学変数を観測する際に付加される観測ノイズである。

力学系 $\dot{X} = F(X)$ に以下のように白色ガウスノイズ $\xi(t)$ を加え、これをシステムノイズと呼ぶ。

$$\dot{X} = F(X(t)) + \xi(t) \tag{2.17}$$

ここで白色ガウスノイズ $\xi(t)$ は $E[\xi(t)\xi(s)] = \rho_S^2 I_n \cdot \delta(t-s)$ を満たす。ただし $E[\cdot]$ は期待値を表し、I_n は n 次元単位行列、ρ_S^2 は連続過程におけるシステムノイズの分散、$\delta(t)$ はディラックのデルタ関数である。なお n はシステムの次元である。(2.17) 式で与えられる $X(t)$ は確率過程になる。時間間隔 δt_R で確率微分

方程式 (2.17) を離散化すると以下のようになる。

$$X(t+\delta t_R) = X(t) + \delta t_R \cdot F_{RK}(X(t)) + \xi_S(t) \tag{2.18}$$

ただし、$F_{RK}(X(t))$ は時間間隔 δt_R の Runge-Kutta 法に基づく真の離散化ベクトル場である。また、離散時間におけるシステムノイズ $\xi_S(t)$ は各時刻で独立に生成され、分散 $(\delta t_R \rho_S^2)$ を持つガウスノイズである。

一方、時刻 t において力学変数 $X(t)$ を観測する際に以下のように付加されるノイズを観測ノイズと呼ぶ。

$$Y(t) = X(t) + \xi_O(t) \tag{2.19}$$

実際の観測値は $Y(t)$ とする。離散時間における観測ノイズ $\xi_O(t)$ は各時刻で独立に生成され、分散 $(\delta t_O \rho_O^2/2)$ を持つガウスノイズであると仮定する。これらのノイズの大きさは、時間間隔 δt_O での観測値による離散化ベクトル場 $(Y(t+\delta t_O) - Y(t))/\delta t_O$ における実効的ノイズの分散がそれぞれ $(\rho_S^2/\delta t_O)$ と $(\rho_O^2/\delta t_O)$ になるように定められている。

ノイズの大きさはアトラクタ上での状態変数 X の分散 $\chi^2 \equiv E[|X - E[X]|^2]/n$ に比例して与える。例えばシステムノイズの分散は $\delta t_R \rho_S^2 = \delta t_R (\kappa\chi)^2$ で与えられる。2.3章以降の説明では κ の百分率を用いてノイズの大きさを表す。

2.3. 部分観測問題

本研究では、力学変数の一部のみが観測される部分観測という状況を想定する。通常、ある現象はほかの現象と相互に影響しあっており、その現象の数の変数を持っていると考えることができる。ある現象を位相空間上で正しく見るためには、その全ての力学変数を観測しそれぞれを軸に取った次元上で観測しなければならない。しかし、実問題において全ての現象を数え上げることは困難であり、また未知のシステムの全ての現象を観測できることは殆んど期待できない。一般的には、その力学系の状態変数の数よりもはるかに少ない観測変数が利用できるのみである。このような部分観測問題では、観測された 1 変数の時系列データから、元の状態空間での位相構造を埋め込みにより構成する。ここで埋め込みとは、1

次元時系列データから元の力学系の同定を可能とするのに十分な次元を確保する技法である。本研究では、2 種類の埋め込み手法について検討した。まず初めにカオス力学系の再構成に一般に用いられる遅れ座標埋め込みを用いた場合の実験結果について述べる。次に、従来の手法に加えてさらに平滑化フィルタを用いた埋め込み手法を提案し、実験結果を報告する。

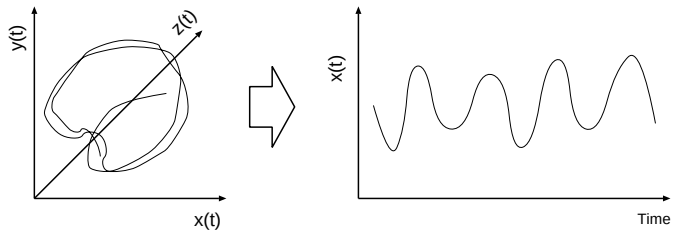

Figure 2.6 部分観測の概念図

2.3.1 遅れ座標埋め込み

ここでは力学変数の一部、変数 x だけが観測されると仮定する。学習器は他の力学変数である y および z を観測することはできない。この実験は佐藤ら [66] によって再帰型ニューラルネットワーク (Recurrent neural network:RNN) に隠れ変数を持つローレンツ方程式を学習させることによって調べられた。それによると、観測変数の軌道がかなりよく再現されたとはいえ、座標変換の不定性ゆえに、RNN で学習された力学系がもとの力学系と等価であるかどうかは確認できないという問題点があった。本研究では、この問題点を避けるために遅れ座標による埋め込み手法を用いる。

変数 $x(t)$ に対する遅れ座標は以下のように定義される。

$$Z(t) \equiv (x(t), x(t-\tau), x(t-2\tau), ..., x(t-(d-1)\tau)) \tag{2.20}$$

ここで d と τ はそれぞれ再構成次元、遅れ時間と呼ばれる。埋め込み定理 [67, 79] によれば、d_S をアトラクタの次元としたとき、再構成次元 d が $2d_S$ より大きい

と以下のような遅れ座標の滑らかなベクトル場が存在する。

$$\dot{Z} = H(Z) \tag{2.21}$$

ベクトル場 H はもとのベクトル場 F の位相的な構造を保存する [67]。

従来の研究 [13] では、遅れ座標 $Z(t)$ から $x(t+\tau)$ を予測するようにトレーニングされることが多かった。このアプローチによると、予測はストロボ的に軌道 $\{x(t+\tau), x(t+2\tau), ...\}$ のみを生成し、それ以外の時刻における予測はできない。

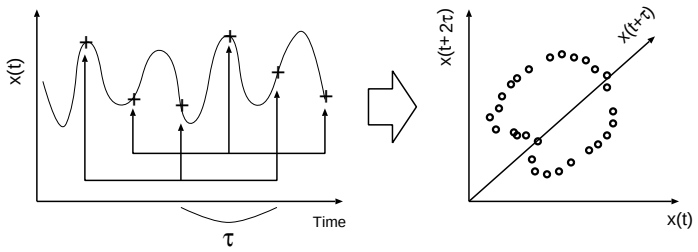

Figure 2.7 遅れ座標埋め込みの概念図

そこで本研究では、遅れ座標空間のベクトル場 $H(Z(t))$ を時間間隔 δt_O で離散化した離散化ベクトル場をターゲットとして、NGnet の学習を行う。遅れ座標空間でのベクトル場はシステムの軌道に沿って以下の大域的制約を持つ。

$$H_{n+1}(Z(t+\tau)) = H_n(Z(t)) \quad n = 1, ..., d-1 \tag{2.22}$$

ただし、$H_n(Z)$ はベクトル場 $H(Z)$ の第 n 成分を表す。学習過程にこの制約を課すためには時間逆伝播法 [62] が必要になり、オンライン EM アルゴリズムが適用できない。そこで、学習時にはこの制約を忘れて学習を行う。

学習後にシステムを動作させる場合、現在の状態 $Z(t)$ における離散化ベクトル場が NGnet の出力として計算される。その後、次の時刻の値 $Z(t+\delta t_O)$ がベクトル場を用いて計算される。この過程において制約条件 (2.22) は無視される。しかし、NGnet がベクトル場を精度良く近似できていれば、学習した NGnet が生成する軌道は制約条件をほぼ満足することが期待される。

2.3.2 積分埋め込み

本節では、連続時間力学系を扱うための平滑化フィルタに基づく新しい埋め込み手法を提案する。

上で述べた遅れ座標埋め込み手法は離散時間システムを扱う上で有効であるが、大域的な制約を持つという困難がある。また多くの場合、実際のデータは連続時間動的システムを離散時間でサンプルした時系列であり、遅れ時間ごとの値のみを用いる上記の手法は、あまり適しているとは言えない。 ここでは、連続時間システムを扱う手法として時系列データの平滑化積分による埋め込みを提案する。この手法は平滑化フィルタを用いるため、よりノイズに強いことが期待される。

ある連続時間 t についての時系列 $x(t)$ の平滑化データは以下の線形微分方程式で得られる。

$$\begin{aligned} y_0(t) &= x(t) \\ \tau \dot{y}_k(t) &= -y_k(t) + y_{k-1}(t) \quad (k \geq 1) \end{aligned} \tag{2.23}$$

ここで、$\dot{y} \equiv dy/dt$ は変数 y の時間微分である。$y_k(t)$ は時刻 t までの y_{k-1} を遅れ時間 τ で平滑化することによって得られる。

$$y_k(t) \equiv \frac{1}{\tau} \int_{-\infty}^{t} e^{-(t-s)/\tau} y_{k-1}(s) ds \tag{2.24}$$

$y_k(t)$ はまた、以下の IIR フィルタを用いて $x(t)$ から求められる。

$$y_k(t) = \int_{-\infty}^{t} G_k(t-s) x(s) ds \tag{2.25}$$

$$G_0(t) = \delta(t) \tag{2.26}$$

$$G_k(t) = \frac{1}{\tau(k-1)!} \left(\frac{t}{\tau}\right)^{k-1} e^{-t/\tau} \quad (k \geq 1) \tag{2.27}$$

ここで $\delta(t)$ は Dirac のデルタ関数である。フィルタ関数 $G_k(t)$ は $t = (k-1)\tau$ にピークを持つ。図 2.8に $\tau = 14$ の平滑化フィルタ関数 $(k = 2, 3, 4, 5, 6)$ の形状を示す。

時系列がサンプリング時間 Δt ごとに観測される、すなわち $\{x(n\Delta t)|n = 0, 1, ...\}$ が観測されるとすると、微分方程式 (2.23) は時間間隔 Δt で離散化され、以下の

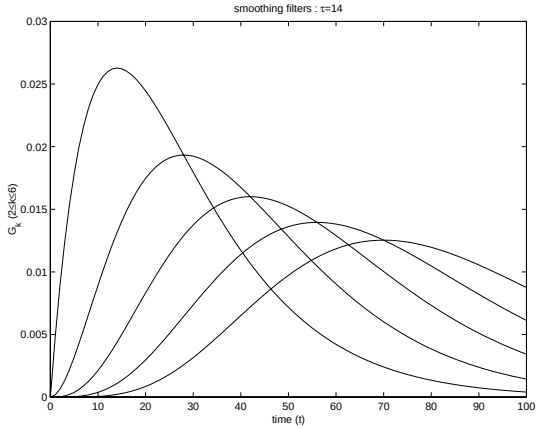

Figure 2.8 積分埋め込み法の平滑化フィルタ

線形差分方程式で書くことができる。

$$\begin{aligned} y_0(n+1) &= x((n+1)\Delta t) \\ y_k(n+1) &= \lambda y_k(n) + (1-\lambda) y_{k-1}(n) \end{aligned} \tag{2.28}$$

ここで λ は減衰係数であり、$\lambda = e^{-\Delta t/\tau}$ である。

式 (2.28) で計算された平滑化データを用いて、新しい座標を以下のように定義する。

$$Y(n) \equiv (y_0(n), y_1(n), \cdots, y_{d-1}(n)) \tag{2.29}$$

d は座標の次元である。k 番目のフィルタ関数は $(k-1)\tau$ にピークをもつため、この座標系は遅れ時間 τ での遅れ座標系と類似している。また、平滑化フィルタ関数がデータに独立に与えられているノイズを除去することが期待される。この新しい座標 (2.29) を積分座標、積分座標への埋め込み手法を積分埋め込みと呼ぶ。

NGnet は、観測時系列 $\{x(n\Delta t)|n = 0, 1, ...\}$ からベクトル場を学習する。つまり離散時刻 n での NGnet への入力は $Y(n)$ であり、ターゲット出力は x の離散化速度、すなわち $(x((n+1)\Delta t) - x(n\Delta t))/\Delta t$ である。

学習後の NGnet が離散化ベクトル場 $F_I(Y)$ を近似することができれば、入力として得られた $\hat{x}(n\Delta t)$ と $\hat{Y}(n)$ から $\hat{x}(n\Delta t) + \Delta t F_I(\hat{Y}(n))$ を計算することによって、次の時刻の観測変数 $\hat{x}((n+1)\Delta t)$ を予測することができる。この $\hat{x}((n+1)\Delta t)$

から次の時刻での入力となる $\hat{Y}(n+1)$ が計算できる。この過程を繰り返すことにより、x の軌道が学習後のシステムによって推定される。

2.4. シミュレーション実験

2.4.1 遅れ座標埋め込みによるレスラーアトラクタの学習

本章では、NGnet 用いて遅れ座標空間に埋め込まれたレスラーアトラクタを学習する。遅れ座標空間への埋め込みは、レスラーシステムの状態変数 x の時系列を用いて行い、NGnet は遅れ座標空間での離散化ベクトル場を近似するように学習を行う。レスラーシステムのパラメータは $d = 3, \tau = 16 \times \delta t_O$ とした。遅れ時間 τ の値は相互情報量 [1] から決めた。ここで 2 つのデータセット x と y の相互情報量は以下の式で計算できる。

$$I(x,y) = \sum_{x,y} P(x,y) \log_2 \left[\frac{P(x,y)}{P(x)P(y)} \right] \tag{2.30}$$

遅れ時間 τ は $x(t)$ と $x(t-2\tau)$ の相互情報量の最初の極小値 [1] よりもやや小さい値として選んだものである。またレスラーアトラクタが $d = 3$ で埋め込めることは global false nearest neighbors 法 [41] を用いて確かめた。すなわち、元のアトラクタの離れた点が埋め込まれたアトラクタにおいても必ず離れていること、つまり埋め込まれたアトラクタが自己交差をしないことを実験的に確認した。図 2.9に遅れ座標空間内に埋め込まれたレスラーアトラクタの相図を示す。遅れ座標系の $x(t+\tau)$ と $x(t+2\tau)$ がそれぞれ状態変数 y と z に直接的に対応しているわけではないが、全体的な形状としては元のアトラクタ (図 2.2) と同様な位相構造を持っていることが分かる。

レスラーシステムの解軌道から求められた遅れ座標系でのデータを 30,000 個学習させた後に NGnet を動作させると、図 2.9のレスラーアトラクタに良く似たカオスアトラクタを生成することができた。学習後の正規化誤差は 0.0093%であり、部分観測という状況においても NGnet は少ないデータ点からカオスアトラクタを良く再現することができると言える。ここで正規化誤差とは、ベクトル場について NGnet の推定値と真値の平均二乗誤差をアトラクタ情の 50,000 点で求め、これをアトラクタ上でのベクトル場の分散で正規化したものである。

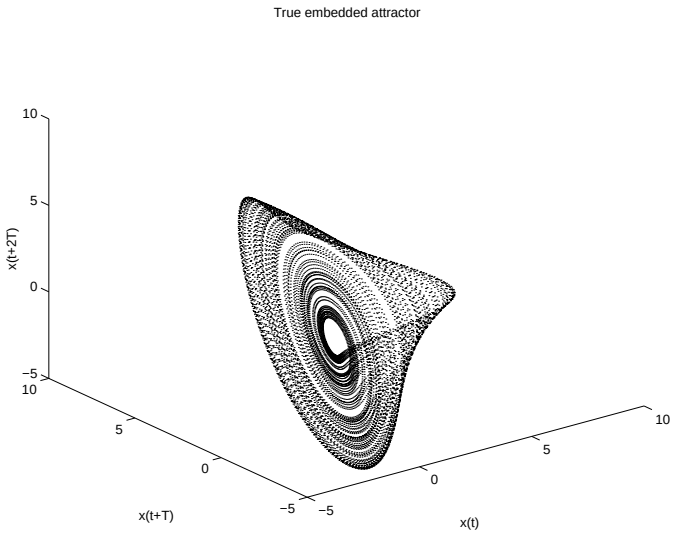

Figure 2.9 遅れ座標空間内に埋め込まれたレスラーアトラクタの相図

v 本手法のノイズへの耐性を調べるため、学習データに 2 種類のノイズを付加し、実験を行った。図 2.10はシステムノイズと観測ノイズをそれぞれ 10%ずつ付加したデータを 2,000,000 個学習した後の NGnet が生成したカオスアトラクタの相図である。図 2.10は元のアトラクタ (図 2.9) と類似しており、NGnet が離散化マップを良く近似できることが分かる。ここで 10%のノイズと言っても、ベクトル場に対しては約 45%の変化を与える大きなノイズであることに注意する。

学習時におけるノイズの影響を定量的に調べるために、学習後の NGnet により生成されたアトラクタの相関次元と最大リアプノフ指数を調べ、真の埋め込みアトラクタの値と比較し、表 2.1にまとめた。この実験ではノイズの大きさによって学習データ数を変えた。これはノイズが小さい場合に必要以上に学習データを与えると生成軌道がリミットサイクルに落ちてしまい、一方ノイズが強い場合では十分な数の学習データから学習を行わないとカオスアトラクタを生成することができないためである。ユニット数の初期値は 200 個、上限は 500 個とした。表 2.1においてユニット数 (NU) とは学習後の NGnet が用いているユニットの個数である。ベクトル場の正規化誤差 (nMSE) は、アトラクタ上 (on att.) とアトラクタ周辺 (around) のそれぞれ 50,000 点の軌道について初期値を変えて 10 セットずつ

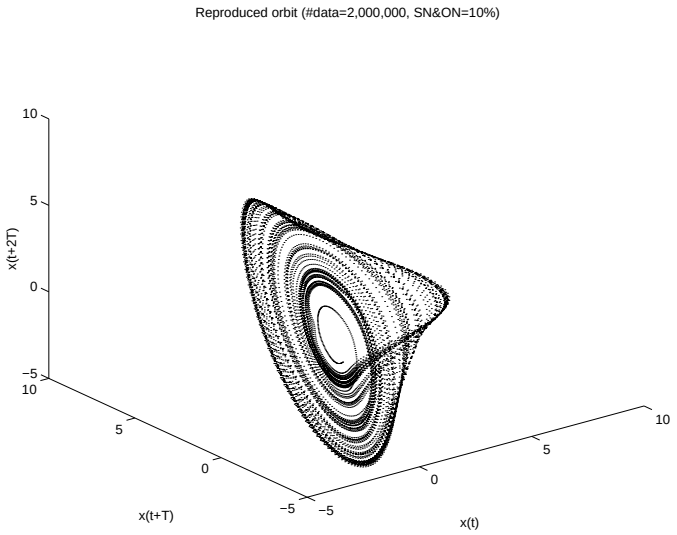

Figure 2.10 学習後の NGnet が生成したカオスアトラクタの相図

調べた結果の平均である。ここでアトラクタ周辺の誤差は、3%のシステムノイズにより挙動が乱されたシステムが生成する軌道上で評価した。相関次元 (CD) は、学習後の NGnet によって生成された軌道上の 50,000 点について初期値を変えて 4 セットずつ調べた際の平均値である。最大リアプノフ指数 (MLE) は、10,000 点の生成軌道に沿った各点でのヤコビ行列を用い、グラム-シュミットの直交化を利用して求めた。ヤコビ行列は離散化された写像関数 $X(t+\delta t_O) = K(X(t))$ を各点の近傍 32 個のデータ点から 2 次の多項式で近似し、これを微分することにより計算した。

学習データにシステムノイズが含まれる場合、ノイズが大きくなるにつれて正規化誤差が増加し、ユニット数が増える傾向にある。しかし、相関次元と最大リアプノフ指数は真の値とかなりよく一致しており、カオスの力学的性質は良く学習できていることが分かる。15%という非常に大きいノイズを加えたときでさえ、NGnet は元の力学系のカオス的性質を良く再現できていると言える。また、ノイズなしデータを学習した場合と比較してみると、アトラクタ上での誤差は大きくなるが、アトラクタ周辺の誤差は小さくなる。このことから、システムノイズを乗せたデータを学習させた方が、かえってアトラクタ近傍の近似精度が高くなっ

Table 2.1 NGnet により生成された遅れ座標空間内でのカオスアトラクタの性質

		NU	nMSE		CD	MLE
			on att.	around		
True embedded					1.835±0.029	0.446±0.005
noiseless		164	0.002	0.067	1.923±0.008	0.369±0.011
system noise	5%	124	0.020	0.023	1.843±0.044	0.406±0.011
	10%	257	0.052	0.056	1.794±0.046	0.324±0.007
	15%	500	0.010	0.105	1.836±0.022	0.404±0.004
observation noise	5%	122	0.044	0.133	1.731±0.018	0.697±0.050
	10%	134	0.141	0.424	1.419±0.066	1.014±0.053

ていることが分かる。一方、データに観測ノイズが含まれる場合、10%までの比較的小さい観測ノイズであれば、NGnet はカオスアトラクタを生成することができる。しかし、再構成されたアトラクタは見た目に痩せており、相関次元が小さくなる。生成時に与える初期値によってはリミットサイクルに落ちる場合もある。これはカオスアトラクタの詳細なフラクタル構造が観測ノイズによって不鮮明になっていることを意味する。これは、学習データ数を増やすことによっても改善されなかった。また、さらにノイズが大きくなると、学習データが十分な個数与えられても NGnet の生成軌道は止まってしまう。

力学システムを同定することによって、観測状態変数 $x(t)$ の将来の振る舞いを予測することができる。図 2.11に学習後の NGnet を用いて予測を行った時の各時刻における予測誤差を示す。図の縦軸は予測誤差、横軸は時間を表す。実線はノイズなしデータを、破線と一点鎖線は、それぞれ 5%のシステムノイズと 5%の観測ノイズを付加したデータを 1,000,000 個学習させた場合の予測誤差の時間変化である。予測誤差は初期値を 100 回変えて調べた結果の平均値であり、各々の初期値は学習データとは別の軌道からランダムに選んだ。ノイズなしデータの場合、予測時間が $t = 30$ になる前、つまり将来の 600 個のデータ点の間は、生成軌道が真の軌道とほぼ一致する。システムノイズが乗ったデータを学習させた場

合、予測時間 $t = 30$ 位までは軌道の良い予測が可能であり、ノイズなしデータの場合と同程度の予測精度を保つことができる。一方、観測ノイズでは予測時間 $t = 10$ 位までは軌道の良い予測が可能である。また、ノイズが大きくなっても予測精度が極端に劣化することはなかった。

カオス力学系では最大リアプノフ指数を MLE とする時、誤差が平均として $\exp(\mathrm{MLE} \cdot t)$ のように指数的に拡大される。この最大リアプノフ指数より定まる予測の時定数 $(1/\mathrm{MLE}) \approx 2.3$ に比べて上記の予測時間は非常に長く、NGnet の予測性能が非常に良いことが分かる。しかしいずれの場合も予測時間が長くなると軌道がずれていく。これは、カオス力学系が誤差を指数的に拡大するという一般的な性質によるものである。

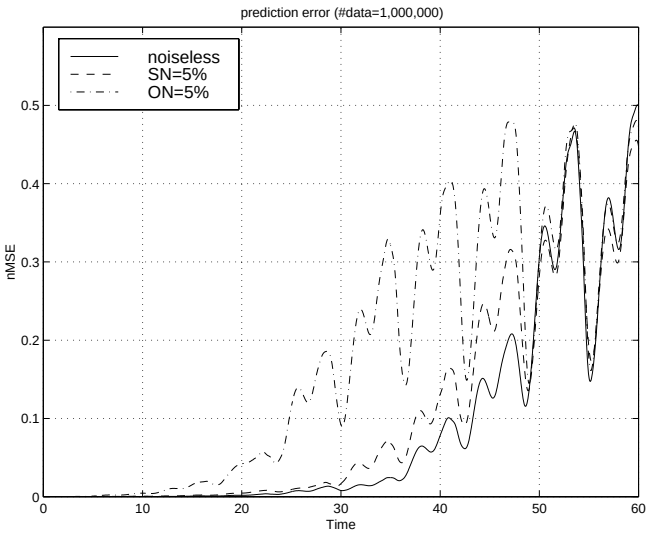

Figure 2.11 部分観測時の予測精度の比較

ところで、(2.10d) 式における σ_i^2 は各ユニットの出力の誤差分散を表しているので、これらをユニットの活性度 $\ll 1 \gg_i$ の重み付けで平均することによりノイズの分散を推定することができる。

$$\left(\sum_{i=1}^{M} \sigma_i^2 \ll 1 \gg_i\right) \Big/ \left(\sum_{i=1}^{M} \ll 1 \gg_i\right) \tag{2.31}$$

Table 2.2 ノイズの大きさの推定

		real(%)	estimate(%)
System noise	5%	3.6	3.2
	10%	14.5	13.0
	15%	32.6	36.3
Observation noise	2%	0.6	0.5
	4%	2.3	1.8

表 2.2に、様々なノイズについて、ベクトル場に与える実質的なノイズの分散 (real) と (2.31) 式による推定分散値 (estimate) を調べた結果をまとめた。これらの結果は、我々の手法によって、データに含まれるノイズの大きさを良く推定できることを示している。

2.4.2 積分埋め込みによるローレンツアトラクタの学習

本節では、提案した積分埋め込み法を用いた実験の結果を報告する。学習対象としてはローレンツ系を用い、遅れ座標埋め込みを用いた場合 [34] と比較する。ここで各パラメータは $d = 3, \tau = 10 \times \Delta t, \Delta t = 0.01$ とした。遅れ時間 τ は y_0 と y_2 の相互情報量の最初の極小値 [1] よりやや小さい値を用いた。図 2.12はこの τ 値を用いた際の積分空間内での $Y(n)$ の相図であり、カオスアトラクタであることが分かる。また、2 枚羽の形状をしており、原点と 2 枚の羽の中心に特異点を持つという、元のローレンツ方程式の軌道と類似した位相的構造を持っている。埋め込みアトラクタの力学的特性を調べたところ、相関次元は 2.033±0.021、最大リアプノフ指数は 1.026±0.003 であった。元のローレンツアトラクタの値は相関次元 2.054±0.006、最大リアプノフ指数 0.958±0.014 であり、積分埋め込み法が元の力学的性質を再現していることが分かる。

積分空間内に埋め込まれたアトラクタの軌道点を用いて学習を行い、学習後の NGnet にアトラクタの再構成をさせた。図 2.13はノイズなしデータを 20,000 点学習した後の NGnet が生成した軌道である。このときの正規化誤差は 0.1%とな

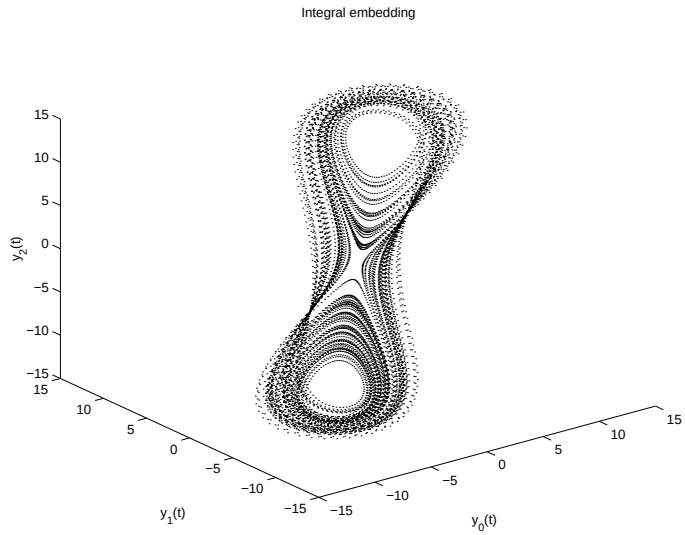

Figure 2.12 積分座標空間内に埋め込まれたローレンツアトラクタの相図

り、少ないデータ点からでもアトラクタの再構成が可能であることが分かる。

また、図 2.14は、ノイズなしデータを 1,000,000 点学習した後の NGnet のユニットの受容野である。図の各楕円は i 番目のユニットの受容野を表し、楕円の軸は共分散行列 Σ_i の第一主成分と第二主成分に対応する。楕円の中心位置はユニットの中心 μ_i である。288 個のユニットがアトラクタ全体を覆うように配置されていることが分かる。

本手法のノイズへの耐性を調べるため、学習データに 2 種類のノイズを付加し、実験を行った。ノイズが大きくなるにつれて、アトラクタ上の正規化誤差が大きくなり、力学系を近似するのに必要な学習データの数が増加する。しかし、十分な数の学習データが与えられると、学習データに非常に大きいノイズが付加されたときでさえ、元のアトラクタと類似したカオスアトラクタを生成することができる。近似精度を定量的に評価するために、学習後のユニット数 (NU)、正規化誤差 (nMSE)、再構成されたアトラクタの相関次元 (CD) と最大リアプノフ指数 (MLE) について調べ、表 2.3にまとめた。学習データ数は 1,000,000 点、ユニットの初期値と最大値は、それぞれ 300 個と 1,000 個とした。学習後のユニット数が初期値より減少しているのは、2.1章の最後で述べたように、あまり使用されて

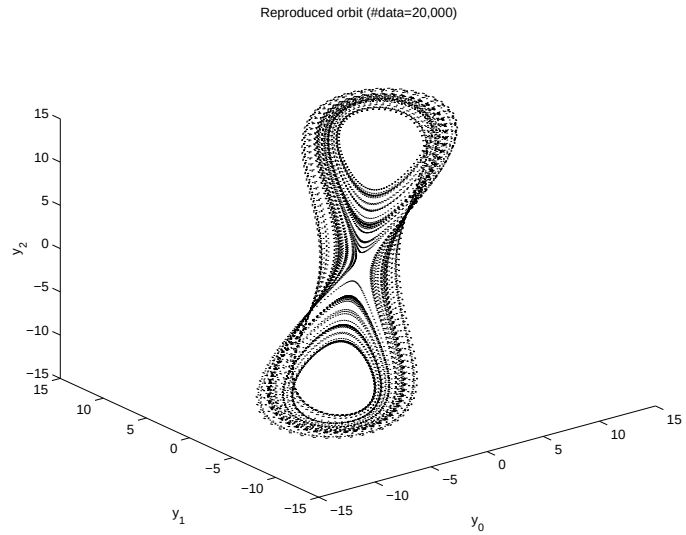

Figure 2.13 再構成された積分埋め込みローレンツアトラクタの相図

いないユニットは削除されるためである。

表 2.3から分かるように、ノイズが大きくなるにしたがって正規化誤差は増大するものの、本手法は非常に大きいノイズ、すなわちシステムノイズが 50%、または観測ノイズが 70%含まれた学習データからでも、カオス力学系の性質を精度良く再現することができる。しかしデータにさらに大きいノイズが含まれると、学習後の NGnet による再構成アトラクタの相関次元が下がり、場合によってはリミットサイクルに落ちてしまう。このノイズが大きくなるにつれてアトラクタの相関次元が下がるという傾向は 2 種類のノイズどちらにも見られるが、観測ノイズを付加した場合の方がややその傾向が強い。これは、観測ノイズを乗せたデータを学習させると空間的な平滑化が行われ、各ユニットが担当する領域が広くなりやすいためだと考えられる。このユニットの領域の拡大が、カオスアトラクタの複雑度の減少を導く。

また、前節で述べた遅れ座標埋め込み手法との比較をするため、遅れ座標埋め込みを用いた場合 [34] の結果を表 2.4に示す。これは 2,000, 000 個の軌道点を学習した後の結果である。ユニットの初期値と最大値は、それぞれ 300 個と 1,000 個とした。遅れ座標埋め込みを用いた場合、システムノイズは 15 %、観測ノイ

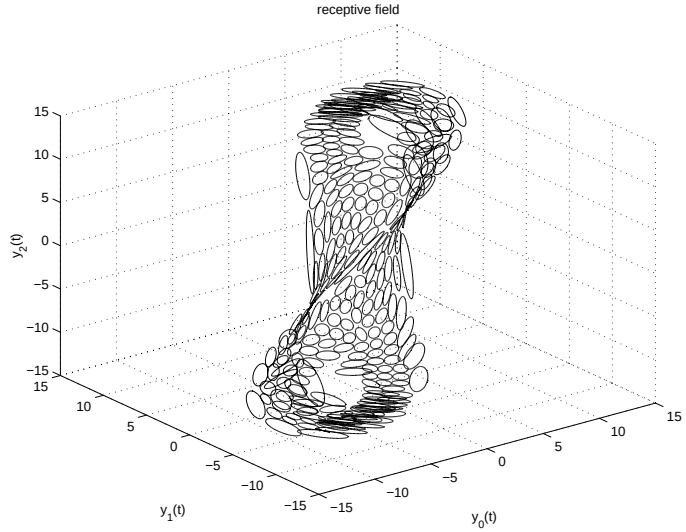

Figure 2.14 学習後の NGnet の受容野

ズは 20%までなら再構成が可能であり、全変数を観測できる場合[1]と比べて極端に耐性が下がる [34]。しかし積分埋め込み手法では、全変数観測時よりもノイズへの耐性が上がり、さらに、再構成されたアトラクタの力学系近似能力も向上することが分かった。また、積分埋め込み手法では必要なデータ数が少ない上、遅れ座標と違ってノイズが大きくなってもユニットの数が増大しないため学習が早い。ユニット数が少ないため正規化誤差が大きくなるが、力学的な性質はよく再現できている。つまり、新しい手法は従来のものよりノイズに強く学習が早いという 2 つのメリットがある。

次に、システムノイズと観測ノイズの学習に与える影響の違いについて考える。システムノイズについては、遅れ座標埋め込み時と同様にノイズが大きくなるに従って相関次元が減少する傾向がある。しかし、50%という非常に大きいシステムノイズが付加されたデータを学習した場合でも、元の力学的性質を良く近似することができる。観測ノイズについては、ノイズへの耐性はシステムノイズの場合より良くなっているが、ノイズが大きくなるにつれて最大リアプノフ指数が大

[1]ローレンツシステムを用いた場合、システムノイズ 40%、観測ノイズ 70%まで再構成が可能であった。

Table 2.3 積分座標空間内でのローレンツアトラクタの力学的性質

		NU	nMSE	CD	MLE
True embedded				2.033±0.021	1.026±0.003
noiseless		288	0.002	2.033±0.023	0.958±0.001
System noise	10%	295	0.015	2.008±0.024	1.047±0.020
	30%	283	0.084	1.935±0.026	0.892±0.016
	50%	249	0.218	1.857±0.031	0.790±0.008
Observation noise	10%	280	0.208	2.006±0.024	1.247±0.005
	30%	290	1.103	1.978±0.025	1.822±0.028
	50%	295	2.820	1.855±0.024	1.898±0.086
	70%	284	5.721	1.847±0.032	2.601±0.089

きくなる。この、観測ノイズが力学系の不安定性を増大させるという傾向は、遅れ座標埋め込みを用いた場合により顕著にみられる。

2 種類のノイズへの耐性を比較すると、いずれの埋め込み手法を用いた場合でも、システムノイズは学習をそれほど阻害せず、場合によってはアトラクタ周辺における汎化能力を高めるように働き、一方で観測ノイズは学習を阻害するように働くことが分かった。この理由として以下のことが考えられる。ノイズのない学習データはアトラクタ上でのベクトル場の正確な情報を持っているが、アトラクタから外れた点における情報は全く持っていない。システムノイズを加えた学習データは軌道が毎時刻ノイズによって乱されるために、アトラクタの近傍におけるノイズを含んだベクトル場の情報を持っている。このためシステムノイズを加えて学習させた場合、真のアトラクタ上でのベクトル場についてはノイズなしの場合に比べて近似精度が悪くなるが、アトラクタの近傍におけるベクトル場の近似精度はかえって良くなり、学習の汎化能力が上がると思われる。ただし、システムノイズを加えた学習データにおけるベクトル場のノイズ成分は非常に大きいので、以上のことを考え合わせてもシステムノイズにより汎化能力が上がるのは、NGnet の学習能力に負うところが大きいと考えられる。2.1章で述べたよう

Table 2.4 遅れ座標空間内でのローレンツアトラクタの力学的性質

		NU	nMSE	CD	MLE
True embedded				2.021±0.007	0.976±0.012
noiseless		387	0.006	1.923±0.063	0.882±0.001
System noise	5%	292	0.006	1.972±0.020	0.968±0.004
	10%	450	0.020	1.940±0.024	1.032±0.004
	15%	813	0.075	1.865±0.034	1.480±0.066
Observation noise	5%	290	0.025	1.857±0.047	2.821±0.199
	10%	504	0.060	1.729±0.050	4.208±0.132
	15%	943	0.146	1.732±0.040	4.250±0.272
	20%	1000	0.120	1.603±0.044	3.997±0.200

に、オンライン EM アルゴリズム [65] では各ユニットのパラメータは統計量の重み付き平均値を用いて表されるので、この平均操作がノイズ除去の効果を持つと考えられる。一方観測ノイズを加えた場合、観測軌道はノイズによって乱されるために、NGnet に対する入力データはアトラクタの近傍で与えられる。しかし、教師データはアトラクタ上でのノイズを含んだベクトル場の情報しか持っていない。このため観測ノイズは学習の汎化能力を高めることが少なく、学習を阻害するように働くことが多いと考えられる。

次に、学習後の NGnet が将来の軌道をどの程度まで予測できるかを調べた。図 2.15および図 2.16 に結果を示す。各図の実線はノイズなしデータを 1,500,000 個学習させた場合の予測誤差の時間変化である。また、図 2.15の破線と一点鎖線は、それぞれ 5%と 15%のシステムノイズを付加したデータを、図 2.16の破線と一点鎖線は、それぞれ 2%と 4%の観測ノイズを付加したデータを 1,500,000 個学習させた場合の予測誤差の時間変化である。図の各線は初期値を 100 回変えて調べた結果の平均を示している。ノイズのないデータを用いた場合、予測時間が $t = 30$ になる前、つまり将来の 600 個のデータ点の間は、平均的に予測軌道が真の軌道とほぼ一致する。5%のシステムノイズが乗ったデータを学習させた場合でも、予

測時間 $t = 20$ 位までは軌道の良い予測が可能である。また、ノイズが 15%と強い場合でも充分な学習データが与えられれば予測精度が劣化することはなかった。観測ノイズに関しては、ノイズが強くなると大幅に予測精度が劣化するが、2%の弱いノイズであればノイズなしの場合と同程度の予測が可能である。これらの予測時間は、最大リアプノフ指数より定まる予測時定数 $(1/\mathrm{MLE}) \approx 2.3$ に比べて非常に長く、NGnet の予測性能が非常に良いことが分かる。しかし、予測時間は初期値に依存し、$t = 0.6$ 程度までしか予測できない場合もある。これは、ローレンツアトラクタの特異点である原点の周りに予測軌道が近づいた際に、真の軌道と反対側の羽に飛ばされることがしばしば起こるためである。

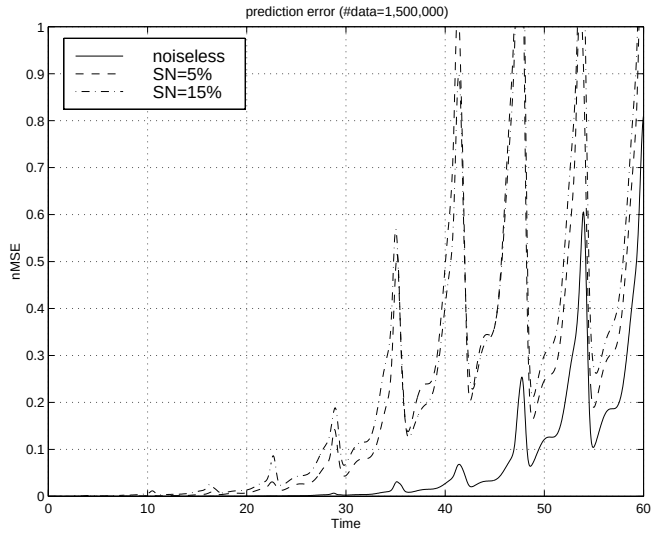

Figure 2.15 学習データにシステムノイズが付加されている場合の予測誤差

2.5. 本章のまとめ

本章では、未知の力学システムから部分的に観測される時系列を用いて、システムを同定する手法について 2 種類の手法を用いて議論した。2 種類とは、部分観測状況を扱うために従来しばしば用いられてきた遅れ座標埋め込みと、本研究で提案する平滑化フィルタを用いた積分埋め込み手法である。対象システムとし

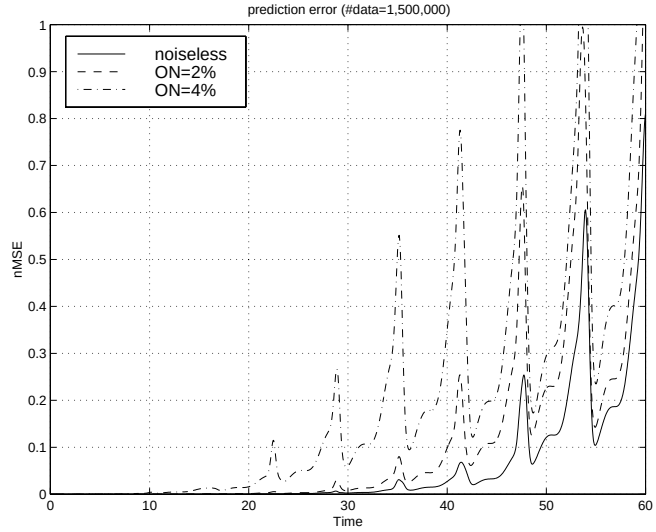

Figure 2.16 学習データに観測ノイズが付加されている場合の予測誤差

ては低次元カオス力学系を用い、NGnet とオンライン EM アルゴリズムを用いて学習を行った。シミュレーション実験により、カオス力学系の 1 変数の解軌道を教師信号として NGnet に学習させることによって、そのダイナミクスの学習が行えることを示した。NGnet の近似精度を定量的に評価するために、再構成されたアトラクタについて、アトラクタの幾何学的複雑度を示す相関次元と、誤差の拡大率の指標となる最大リアプノフ指数を調べた。学習後の NGnet が生成するアトラクタと真のアトラクタの値を比較したところほぼ一致し、NGnet は元の力学系の力学的性質を良く再現できることが分かった。

学習時におけるノイズの影響を調べるために、2 種類のノイズについて検討した。ここで 2 種類とはシステムのダイナミクスそのものを乱すシステムノイズと、システムの力学変数を観測する際に付加される観測ノイズである。ノイズを付加した学習データを用いて実験を行ったところ、大きいノイズが乗ったデータからでも軌道を生成することができ、本手法がノイズに対して非常にロバストであることが分かった。特に、新しい手法である平滑化埋め込み法を用いると、データにさらに大きいノイズが含まれていても同定精度が劣化しないことを確認した。また、それにより大きいノイズを含むデータからの予測精度も従来の手法による

ものよりも大きく向上した。これは、平滑化フィルタが学習データに付加されたノイズを除去する効果を持つためだと考えられる。また、相関次元と最大リアプノフ指数が真の値とより良く一致するため、積分埋め込み手法の近似精度が高いことが分かる。さらに、この手法は必要な学習データ数が少なく、ノイズが大きくなってもユニットの数が増加しないため、従来の手法より学習が早いという長所を持つ。また、力学システムを同定することにより時系列の予測が行えることも示した。NGnet の予測時間は力学系の性質に基づく予測時定数よりも長く、本手法が非常に良い予測性能を持つことが分かった。

Chapter 3

部分観測問題のモデル同定強化学習

本章では、直接観測できない状態を含むような環境を、観測状態の観測経験に基づきシステム同定し、同定された環境モデルを用いてエージェントの意思決定を行う数理モデルとして、モデル同定強化学習を考える。環境モデルの学習、すなわち状態遷移確率の近似は無情報事前知識を用いたベイズ推定によって行う。また、過去の経験に基づく推定の忘却効果を導入する。また、エージェントの行動選択におけるランダム性の制御を行うために、逆温度の制御機構を導入した。この制御は、行動価値関数のばらつきと環境変化の認識の両方に基づいて行う。本手法をバリアのある 2 次元迷路探索問題に適用し、逆温度の制御を行わない手法と比較した。実験により、逆温度を状態に応じて制御することによって、環境の変動にうまく適応できることが分かった。

3.1. 強化学習

強化学習 [78] とは、行動の結果、あるいは状況の良し悪しを表すスカラー信号を元に、試行錯誤を通して、環境に適応するための機械学習の枠組である。強化学習の例題として、迷路の経路探索問題やロボットの自動制御などが挙げられる。本章では、動的な環境すなわち時間と共に変化する環境における強化学習スキーマについて議論する。

離散時間環境における強化学習問題は以下のように定式化できる。時刻 t で環

境の状態 $s(t)$ を観測すると、エージェントはある方策 π に基づいて行動 $a(t)$ を決定する。すなわち、$a(t) = \pi(s(t))$ である。選択された $a(t)$ と環境のダイナミクスにしたがって、状態は $s(t+1)$ に遷移する。この時、状態の良し悪しに応じてエージェントには即時報酬と呼ばれるスカラー値 $r(t+1) \equiv r(s(t), a(t))$ が与えられる。強化学習の目的は、できるだけ多くの報酬を獲得できるように行動を決定することであり、そのためには、将来の報酬を予測することが重要である。なぜなら、目先の即時報酬 $r(t+1)$ を最大化しても、将来の報酬 $r(t+2), r(t+3), \cdots$ が小さければ、結果として得られる報酬の累積は小さくなってしまうからである。そのため、強化学習は「即時報酬の時間和を最大化する方策を獲得する問題」として扱うと都合が良い。即時報酬の時間和としてまず考えられるのは、各時刻での報酬を全て足していく以下の単純和である。

$$R(t) \equiv \sum_{\tau=0}^{T-t} r(t+1+\tau) \tag{3.1}$$

ここで、T は最終時刻を表している。例えばロボットの制御問題のように $T \to \infty$ の問題を扱う場合、(3.1) 式の $R(t)$ は発散してしまうという問題がある。これを避けるために、将来の報酬を徐々に減衰させていく以下の減衰和がしばしば用いられる。

$$R(t) \equiv \sum_{\tau=0}^{\infty} \gamma^{\tau} r(t+1+\tau) \tag{3.2}$$

ここで、γ $(0 \leq \gamma \leq 1)$ は割引率と呼ばれるパラメータであり、$\gamma = 1$ の時、減衰和 (3.2) 式と単純和 (3.1) 式は等価になる。

強化学習では、報酬を評価してその評価を最大化することで学習を行う。現在の状態の良さを表す関数として、状態価値関数を定義する。状態遷移系列はエージェントの行動履歴に依存するため、報酬の時間減衰和 $R(t)$ はエージェントの方策に依存する。そこで、初期状態を s、エージェントの方策を π に固定した時の $R(t)$ の期待値を状態価値関数と定義する。すなわち、状態価値関数 $V(s)$ は以下で定義される。

$$V(s) \equiv E\left[\sum_{\tau=0}^{\infty} \gamma^{\tau} r(\tau+1) \middle| s = s(0)\right] \tag{3.3}$$

状態価値関数 $V(s)$ は、方策 π を固定した場合の状態 s の良さを表すと同時に、初期状態 s を固定した場合の方策 π の良さを表している。

強化学習スキーマは一般に、確率的ではあるが静的な環境における意思決定問題や最適操作問題である、マルコフ決定過程 (Markov decision process: MDP) として定式化される。本研究では、直接観測できない状態変数を持つ部分観測環境を仮定し、部分観測マルコフ決定過程 (partially-observable MDP: POMDP) による定式化を用いる。

3.1.1 マルコフ決定過程

環境がマルコフ的であるとする。すなわち状態 s で行動 a をとり、状態 s' に遷移する確率を $P(s'|s,a)$ とする。状態遷移確率 $P(s'|s,a)$ が分かっている場合、状態 s における状態価値関数 $V(s)$ は以下の最適ベルマン方程式を満たす必要がある。

$$V(s) = \max_a Q(s,a) \tag{3.4a}$$

$$Q(s,a) \equiv r(s,a) + \gamma \sum_{s'} P(s'|s,a)V(s') \tag{3.4b}$$

ここで $Q(s,a)$ は行動価値関数と呼ばれる。$r(s,a)$ は、状態と行動の対に対する即時報酬であり、決定論的であると仮定する。$0 \leq \gamma \leq 1$ は減衰定数である。状態価値関数 $V(s)$ は、将来にわたる減衰重み付き累積報酬和を表す。行動価値関数 $Q(s,a)$ は、状態 s において行動 a を取り、以後は最適な行動選択を行う際に得られる期待報酬和を表す。

$Q(s,a)$ を最大化する方策を最適方策と呼ぶ。マルコフ決定過程の環境下での強化学習の目的は、この最適方策を求めることである。環境のダイナミクスを表す状態遷移確率 $P(s'|s,a)$ が既知の場合、この問題は動的計画法で解くことができる。しかし、強化学習で扱う多くの問題において状態遷移確率は未知である。環境ダイナミクスのモデルを用いずに、経験から直接価値関数を学習するモデルフリー強化学習の 1 つに、TD(Temporal difference) 学習 [76] がある。TD 学習を用いた手法としては、Actor-critic 学習 [5] や Q 学習 [89] などが挙げられる。TD 学習は、以下の TD 誤差を用いて学習を行なう。

$$\delta = (r(s,a) + \gamma V(s')) - V(s) \tag{3.5}$$

1 つ先の状態 s' の価値と現在の状態行動対 (s,a) に対する即時報酬の和である第一項は、実際の状態遷移に基づいた現在の状態行動対に対する価値を表す。この第一項と、現在の予測に基づいた状態 s の価値を表す第二項の差、つまり報酬の予測誤差を TD 誤差という。TD 学習は、この TD 誤差 δ を確率的に小さくする (確率近似法と呼ばれる) ように価値関数を学習する [76]。したがって、状態遷移確率は間接的に学習されることになる。

一方モデル同定強化学習 [77, 48, 18, 21] では、観測された状態変化をもとに、環境のモデル、すなわち状態遷移確率 $P(s'|s,a)$ を近似推定する。そのため、部分観測環境や時間変化する環境といった、複雑な環境を扱うのに適している。また、マルチエージェント系などへの拡張も比較的容易である [45]。モデル同定強化学習では、価値関数の学習と環境モデルの学習は同時に、しかし独立に行われる。

3.1.2 部分観測マルコフ決定過程

本研究では、環境変数のうち一部のみが観測される状況下での最適意思決定問題を考える。このような問題は、部分観測マルコフ決定過程 (POMDP) [39] として定式化することができる。典型的な POMDP は、直接観測できない状態変数 (隠れ変数) を持つマルコフ環境を取り扱う。以後は、環境の状態を $s \equiv (y,z)$ とする。y と z はそれぞれ、観測状態と不観測状態を示す。このモデルでは、真の状態 s に対してマルコフ性を想定する一方で、観測される状態 y は不観測状態が存在するためマルコフ性を保持していない。POMDP を扱う手法を概観すると、無記憶な手法と記憶に基づく手法がある。不観測状態を無視して強化学習を適用する無記憶な手法 [73] では、不観測状態を無視して不完全な観測状態のみで学習を行うが、学習が非常に遅くなるという欠点がある。一方記憶に基づく方法として、信念状態 (belief state)MDP と呼ばれる手法がある。信念状態とは、真の状態の推定値を真の状態空間上の確率分布という形で表現したものである。

信念状態 MDP では、最適ベルマン方程式は以下のように与えられる。

$$V(b) = \max_a Q(b,a) \tag{3.6a}$$

$$Q(b,a) \equiv r(b,a) + \gamma \sum_{b'} P(b'|b,a)V(b') \tag{3.6b}$$

MDP でのベルマン方程式との違いは、状態 s が信念状態 b に置き換えられてい

る点である。信念状態は、状態の確率分布によって表現される。観測状態に対する確率因子は存在しないため、$b = [y, \hat{P}(z)]$ となる。$\hat{P}(z)$ は不観測状態変数の確率分布の推定値である。状態推定器 (SE) が、行動 a によって得られた観測値 y' を用いて、新しい信念状態 b' を推定できる、つまり $SE(b, a, y') \equiv b' = [y', \hat{P}'(z)]$ であると仮定する。ここで、不観測状態の確率分布 $\hat{P}(z)$ は、新しい観測後に変化するかも知れないことに注意する。さらに、報酬関数が不観測状態に依存せずに決まると仮定すると、(3.6) 式は以下のようになる。

$$V([y, \hat{P}(z)]) = \max_a Q([y, \hat{P}(z)], a) \tag{3.7a}$$

$$Q([y, \hat{P}(z)], a) = r(y, a) + \gamma \sum_{y'} P(y'|[y, \hat{P}(z)], a) V([y', \hat{P}'(z)]) \tag{3.7b}$$

本研究では、状態空間と行動空間が離散で有限であるという有限世界を仮定する。このような有限世界においても、不観測変数の確率分布を取り扱うことが困難であるため、(3.7) 式の信念状態 MDP を解くことは難しい。そこで、近似が必要になる。強化学習エージェントが不観測変数の推定値に確信を持っている場合、$[y, \hat{P}(z)]$ は $[y, \hat{z}]$ と等価になる。ここで $\hat{z}$ は z の最尤値である。この近似によって、(3.7) 式は以下のように書ける。

$$V([y, \hat{z}]) = \max_a Q([y, \hat{z}], a) \tag{3.8a}$$

$$Q([y, \hat{z}], a) = r(y, a) + \gamma \sum_{y'} P(y'|[y, \hat{P}(z)], a) V([y', \hat{z}']) \tag{3.8b}$$

しかし、不観測変数の推定値に対する確信度が低い場合、この近似は適切でなく、近似的なベルマン方程式 (3.8) に基づいた方略は真に最適な方略ではなくなる。この問題を解決するために本研究では、3.3.4 節で提案する探索ボーナスと呼ばれる機構を用いる。

3.2. 環境のモデル化

ここでは、不観測多項変数 z を持つ環境を仮定し、その分布 $\hat{P}(z)$ をベイズ推定によって求める。

例として、非常に単純な迷路問題を図 3.1に示す。エージェントの目的は、スタート位置 (S) からゴール位置 (G) まで到達することである。迷路には点線で示

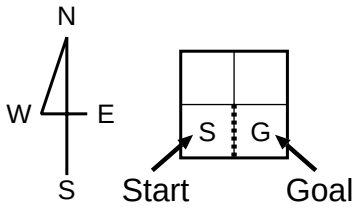

Figure 3.1 単純な迷路の例

されるようなバリアが存在し、エージェントには観測できないものとする。バリアの存在は、バリアを越えようとする行動の失敗によってのみ認識できる。バリアの有無を確率的事象とみなすと、これは二項事象となる。図 3.1の迷路で、バリアがなければ最短距離は 1 ステップだが、バリアが存在すると 3 ステップになる。このような問題は、2 つの方法で定式化することができる。1 つは、バリアの有無を環境の確率的事象であると仮定する、確率的 MDP である。もう 1 つは、バリアの有無を隠れ変数だとみなし、全ての変数が観測できれば環境が決定論的に決まると仮定する、決定論的 POMDP である。

3.2.1 多項モデルのベイズ推定

M 個の可能な値を取る不観測変数 z を仮定すると、z は M 次元ベクトル $z_i \in \{0,1\}$ $(i = 1, ..., M)$ で表現され、$\sum_{i=1}^{M} z_i = 1$ となる。$z_i = 1$ は z が i 番目の値を取ることを示す。多項分布の確率モデルパラメータを $g \equiv (g_1, ..., g_M)$ とすると、$\sum_{i=1}^{M} g_i = 1$ である。迷路問題の例では $(M = 2)$、$z_1 = 1$ と $z_2 = 1$ はバリアの有無をそれぞれ示し、g_1 と g_2 はそれらの確率を示すことになる。T 個の事象系列 $Z \equiv \{z(t)|t = 1, ..., T\}$ を観測した後の尤度は、以下のように与えられる。

$$P(Z|g) = \prod_{t=1}^{T} \prod_{i=1}^{M} g_i^{z_i(t)} = \exp\left(T \sum_{i=1}^{M} \langle z_i \rangle_D \log g_i \right) \tag{3.9}$$

ここで、$\langle z_i \rangle_D \equiv \frac{1}{T} \sum_{t=1}^{T} z_i(t)$ である。(3.9) 式から、多項変数の尤度は指数型であり、その十分統計量は $\langle z_j \rangle_D$、自然パラメータは $\log g_j$ であることが分かる。

ベイズ推定の目的は、観測データ Z からパラメータの事後分布 $P(g|Z)$ を推定

することである。ベイズの定理から、パラメータの事後分布は以下で与えられる。

$$P(g|Z) = \frac{P(Z|g)P(g)}{P(Z)} \tag{3.10}$$

ここで、$P(g)$ はパラメータの事前分布、$P(Z|g)$ はデータに対するパラメータの尤度である。正規化項 $P(Z) \equiv \int P(Z|g)P(g)dg$ はモデル周辺化尤度と呼ばれる。ベイズ推定を行う手法として、試験事後分布 $Q(g)$ を用いて真の事後分布 $P(g|Z)$ を近似する方法がある。すなわち、以下で定義される試験事後分布に関する変分自由エネルギーを最大化することによって実現される。

$$F(Q) = \int Q(g) \log \frac{P(Z|g)P(g)}{Q(g)} dg \tag{3.11}$$

自由エネルギー $F(Q)$ を Q に関して最大化すると、試験事後分布は真の事後分布に等しくなり、これらの分布の KL 距離は最小値 0 になる。(3.11) 式の最大化問題は、変分条件 $\delta F/\delta Q$ を用いて解くことができる。

パラメータ g の事後分布を規定するために自然共役型事後分布を仮定すると、事後分布はハイパーパラメータ $\nu \equiv (\nu_1, \cdots, \nu_M)$ を持つ Dirichlet 分布となる。

$$Q(g|\nu) = \exp\left(\sum_{i=1}^{M} \nu_i \log g_i - \Phi(\nu)\right) \tag{3.12}$$

ここで、$\Phi(\nu)$ は正規化項である。

事前分布として無情報事前分布を仮定すると、尤度の形から、

$$\nu_i = T\langle z_i \rangle_D \tag{3.13}$$

と推定されることになる。Dirichlet 事後分布に関するパラメータの期待値は以下のようになる。

$$\bar{g}_i \equiv \int g_i Q(g|\nu) dg = \frac{\nu_i + 1}{\sum_{j=1}^{M} \nu_j + M} \tag{3.14}$$

すなわち、

$$\bar{g}_i = \frac{T\langle z_i \rangle_D + 1}{T + M} \tag{3.15}$$

となる。これから、$\hat{P}(s_j|s_i, a)$ は $\bar{g}_j$ で推定する。この式の意味することは興味深い。状態 s_i において行動 a をとったとき、状態 s_j に遷移する確率の推定値

$\hat{P}(s_j|s_i,a)$ は、遷移を観測した数 T が多い場合、ほぼ $\langle z_j \rangle_D$、すなわち観測データ中で s_j に遷移した割合になる。一方、観測データ数 T が少ない場合、g_j の推定値は、$1/M$、すなわち可能な行き先に関して全くランダムであることになる。このように、経験によって得られた知識の量を反映した結果が得られる。ここで T は状態 s_i において行動 a を行なった回数であることに注意する。

(3.8) 式では、$\hat{P}(z)$ と $\hat{z}$ を推定する必要がある。(3.15) 式を用いて、$\hat{P}(z_i = 1)$ は $\bar{g}_j (i = 1, \cdots, M)$、$\hat{z}$ は $z_k = 1 (k = \arg\max_i \bar{g}_i)$ と推定する。

3.2.2 オンライン学習と忘却効果

環境が時間とともに変化する動的な環境では、過去の観測に基づく推定は正しくなくなることがあるため、最近の観測を強調して推定を行うべきである。このような推定は、以下のような重み付き変分自由エネルギー [66] を定義することによって実行できる。

$$F(Q|\tau) \equiv \tau\eta(T) \sum_{t=1}^{T} \left(\prod_{u=t+1}^{T} \lambda(u) \right) \int Q(g) \log P(z(t)|g) dg + \int Q(g) \log \frac{P(g)}{Q(g)} dg \tag{3.16}$$

ここで、$\eta(T) = \left[\sum_{t=1}^{T} \left(\prod_{u=t+1}^{T} \lambda(u) \right) \right]^{-1}$ は正規化項である。時間依存の減衰係数 $\lambda(t)$ $(0 \leq \lambda(t) \leq 1)$ は、例えば $1 - \lambda(t) \sim 1/t$ のように、t が増加するにしたがって 1 に近付くようにスケジューリングされる。時間が経つにしたがって、データに対する重みが小さくなっていくため、(3.16) 式によって、最近のデータを強調することができる。言い替えれば、古いデータに対する忘却効果を導入していることになる。この重み付き変分自由エネルギーはオンラインベイズ推定を導く。また、パラメータ τ は、重み付き自由エネルギーの実効的データ数と対応する。もし τ が T より小さいと、新しい自由エネルギー (3.16) は、元の自由エネルギーよりも事前知識を重視したものとなる。つまり、パラメータ τ は、尤度と事前知識のバランスを制御している。本手法でのベイズ推定には無情報事前知識を用いているため、パラメータ τ の減少は不観測変数のランダム推定をすることになる。つまり、このパラメータは環境のダイナミクスに関する知識を忘却することに相当する。

上で述べた考案を元に、本研究では以下のベイズ推定を用いる。

$$\nu_i = \tau\langle z_i\rangle(t) \tag{3.17a}$$

$$\tau^{new} := \begin{cases} \tau^{old}+1 & \text{(after one perception of variable } z\text{)} \\ \kappa\cdot\tau^{old} & \text{(after an episode)} \end{cases} \tag{3.17b}$$

ここで、t 番目の経験後の十分統計量 $\langle z_i\rangle(t)$ はオンライン的に求められる [66]。

$$\langle z_i\rangle(t) = (1-\eta(t))\langle z_i\rangle(t-1)+\eta(t)z_i(t) \tag{3.18a}$$

$$\eta(t) = (1+\lambda(t)/\eta(t-1))^{-1} \tag{3.18b}$$

実効的データ数 τ は、各不観測変数が認識される度に増加する。一方、全ての不観測変数に対する τ の値は、1 回のエピソード毎に減衰係数 $0<\kappa\leq 1$ によって指数的に減少する。ここでエピソードとは、エージェントが 1 回のタスクを遂行するまでの状態遷移系列をいう。

以上に基づいて、状態推定器は以下のように推定することになる。

$$\hat{P}(z_i=1)=\bar{g}_i=\frac{\tau\langle z_i\rangle+1}{\tau+M} \tag{3.19}$$

ここで、$\langle z_i\rangle$ は現在の十分統計量である。$\hat{z}$ は、$k=\arg\max_i \bar{g}_i$ である $z_k=1$ として推定される。ある不観測変数がしばしば認識されると、その実効的データ数が大きくなり、(3.19) 式による推定値は最尤値とほぼ一致する。逆に、不観測変数 z があまり認識されない場合は、その実効的データ数が小さくなる。この場合、エージェントがその変数に対する最近の知識を持っていないため、推定がランダムになる。迷路の例では、エージェントがバリアを越えようとする行動を取らないと、バリアの有無に対する確信度が低くなる。エージェントは、不観測変数、すなわち環境が時定数 $1/(1-\kappa)$ で変化すると推測する。

本研究ではベイズ推定に無情報事前知識を用いたが、これに変わって情報事前知識を用いることも可能である。Dayan ら [18] は、エージェントの探索を促進するために、バリアがなくなりやすいという事前知識を用いている。

3.2.3 状態遷移の推定

迷路探索問題では、不観測変数の確率が状態遷移確率と等価となる。したがって、上記の不観測変数のベイズ推定は、状態遷移確率の推定に用いることがで

きる。

S と A をそれぞれ、状態と行動の集合とする。状態 $s_j \in S$ から行動 $a \in A$ によって状態 $s_i \in S$ に遷移する確率を $P(s_i|s_j, a)$ とする。ある状態行動対 (s_j, a) から到達可能な M 個の状態が存在するとすると、多項変数 z は M 次元ベクトルとして表現される。状態 $s_i \in S$ への遷移が起こると、$z_i = 1, z_k = 0(k \neq i)$ である。パラメータ g_i は、多項モデルの確率モデルを定義する。

3.3.2 節で述べたように、状態遷移確率 $P(s_i|s_j, a)$ は (3.19) 式の $\bar{g}_i$ と推定される。すなわち $\hat{P}(s_i|s_j, a) = \bar{g}_i$ である。実効的データ数 τ は、以下のように更新される。

$$\tau^{new} := \begin{cases} \tau^{old} + 1 & (\text{action } a \text{ is selected at state } s_j) \\ \kappa \cdot \tau^{old} & (\text{after an episode}) \end{cases} \tag{3.20}$$

これは、実効的データ数 τ が増加すると、$\hat{P}(s_i|s_j, a)$ が最尤推定値に近付くことを意味する。一方、実効的データ数が小さいと、推定は $1/M$ に近付き、全ての可能な遷移先についてランダムになる。このように、推定はエージェントの持つ情報量に影響を受ける。τ は、全ての状態行動対 (s_j, a) に対して個々に決められる。

3.2.4 モデル同定強化学習

3.2.3節で述べたように、迷路探索問題では、状態遷移の推定値は不観測変数の推定値と等価になる。

本研究で用いるモデル同定強化学習では、(3.8) 式の行動価値関数は以下で置き換えられる。

$$Q([y, \hat{z}], a) = r(y, a) + \gamma \sum_{y'} \hat{P}(y'|y, a) V([y', \hat{z}']) \tag{3.21}$$

$\hat{P}(y'|y, a)$ は 3.2.3節で述べた手法を用いて求められ、これは、決定論的 POMDP における不観測変数のベイズ推定と等価になる。一方、確率的 MDP のモデル同定強化学習では、ベイズ推定に基づいた確率的環境モデルを推定する。

モデル同定強化学習は、より一般的な問題、すなわち不観測変数と確率的な状態遷移を持つ確率的 POMDP に適用することもできる。このような状況では、不観測変数の存在を環境の確率的事象と見なして、観測状態変数 y に対する状態遷移 $\hat{P}(y'|y, a)$ をベイズ推定によって推定する。

3.3. 行動選択時のランダム性の制御

エージェントが正しい環境ダイナミクスと最適価値関数を獲得できていれば、最適方策は各状態で価値関数を最大化する貪欲な行動を選ぶことになる。このように環境の推定と予測が正確である場合、良い方策とは現在の価値関数に基づく最適行動を選択するものであり、これを exploitation と呼ぶ。しかし、試行錯誤の段階では、エージェントは正しい最適価値関数を知らない。特に部分観測マルコフ決定過程においては、不観測状態変数の推定が不確実であるため、価値関数の近似値が最適価値関数と離れてしまう可能性がある。このような状況において、貪欲な行動は必ずしも最適ではない。また、環境が時間と共に変動する場合、過去の経験によって求められた価値関数は最適ではなくなる。このような状況で最適価値関数を獲得するために、エージェントは現在の価値関数のもとでは最適でないような行動も試みる必要がある。これを exploration と呼ぶ。これらの 2 つの戦略、exploitation と exploration は同時に実行できないため、この 2 つをバランスすることは制御の分野で重要な問題とされてきた [25]。

exploration の手法は、間接探索法と直接探索法の大きく 2 つに分類することができる [82]。間接探索法は、全ての可能な行動に正の確率を割り当てることによって、状態行動空間を全体的に探索しようとする手法であり、semi-uniform(ϵ-greedy) exploration やボルツマン探索法 [78] などがある。一方で直接探索法は、過去の経験から得られた統計値を用いて効率良く探索を行おうとする手法であり、exploration ボーナスがしばしば用いられる。

exploitation と exploration のバランス問題に対し、行動選択時のランダム性を制御することにより、状況に応じて適切に行動を行う手法を提案する。本節では $[y, \hat{z}]$ は s と記述する。

3.3.1 逆温度メタパラメータ

本節では、状態 s で行動 a が確率 $P^\pi(a|s)$ に従って確率的に選択されるものとし、π を確率的方策と定義する。ここで $\int P^\pi(a|s)da = 1$ である。期待報酬量を大きくする方策を決定するためには、現在の行動価値関数に基づいて、

$$\int Q(s,a)P^\pi(a|s)da \tag{3.22}$$

を最大化する確率分布 $P^\pi(a|s)$ を求めることになる。行動数が有限である場合、(3.22) 式を最大化する $P^\pi(a|s)$ は、1 つあるいは数個の行動を除く他の行動に対して確率 0 を割り振るものになる。そのため、環境の変動に追随しにくく、また現在の局所最適方策を改善して良い方策を発見することが困難になる。この問題は exploitation-exploration ジレンマの 1 つである。これを解決するための手法として、確率 $P^\pi(a|s)$ のエントロピーを考慮した以下の自由エネルギーを最大化するようにする。

$$J(P^\pi) = \int Q(s,a)P^\pi(a|s)da - \frac{1}{\beta}\int P^\pi(a|s)\log P^\pi(a|s)da \tag{3.23}$$

(3.23) 式の第一項と第二項はそれぞれ、エネルギー項とエントロピー項と呼ばれる。β は逆温度と呼ばれるパラメータである。$\beta \to \infty$ の極限ではエントロピー項が 0 となり、exploitation を促進させる。一方で、β が小さくなるにつれて自由エネルギーに対するエントロピー項の割合が大きくなり、exploration の効果が強められて探索的な方策を取るようになる。したがって、(3.23) 式の逆温度 β は exploitation と exploration をバランスさせるメタパラメータとなっている。

分布条件 $\int P^\pi(a|s)da = 1$ の下で、(3.23) 式の自由エネルギーを最大化する分布は、変分法を用いて、

$$P^\pi(a|s) = \frac{\exp(\beta Q(s,a))}{\int \exp(\beta Q(s,a))da} \tag{3.24}$$

と求められる [33]。これは soft-max 方策あるいは Boltzmann 選択則 [78] と呼ばれる確率的方策である。温度が高い場合には、行動価値の値に関わらず全ての行動が同程度に起こり、逆に低い場合には、行動価値の値に従って行動の選択確率の差が大きくなるように設定される。Boltzmann 選択則は最適方策ではないので、これを用いた学習は厳密には value iteration にはならず、policy iteration の一種である。しかし、逆温度パラメータ β を徐々に大きくしていくアニーリング法を用いると方策が最適に近付くため、value iteration アルゴリズムが収束するのと類似の議論により収束することが示される。

3.3.2 ランダム性の局所的制御

(3.23) 式は状態 s が与えられた場合の最適化問題として定式化されていた。逆温度メタパラメータが状態に関わらずマクロなパラメータとすることに意義は少

ない。むしろ状態ごとに可変とし、その状態における行動の確信度を表現するという方が尤もらしい。状態に依存したメタパラメータ $\beta(s)$ の制御がメタ制御 [33] である。

逆温度メタパラメータを状態に関わらず一定の値を取るように設定することは、v エネルギー項に対するエントロピーを一定にすることになる。それに対し、エネルギー項は行動価値関数の形状に依存する。例えば、ある状態 s の行動価値関数が全ての行動 a に関して一定である場合、β が大きい値であっても soft-max 方策はランダム方策になる。一方で、行動価値関数が各行動に対して大きくばらついている場合、soft-max 方策は貪欲な行動を取り易くなる。つまり、方策のランダム性は可能な行動に関する行動価値関数のばらつきに依存して決まる。

行動価値関数のばらつきを考慮した正規化 soft-max 方策を以下のように定義する。

$$P^{\pi}(a|s) = \frac{\exp(\beta_0 \tilde{Q}(s,a))}{\int \exp(\beta_0 \tilde{Q}(s,a))da} \tag{3.25a}$$

$$\tilde{Q}(s,a) \equiv \frac{Q(s,a) - E[Q(s,a)]}{\sqrt{E[Q(s,a)^2] - (E[Q(s,a)])^2}} \tag{3.25b}$$

ここで、β_0 は新しい逆温度メタパラメータで定数とする。$E[\cdot]$ は現在の方策に関する期待値を表し、実際の経験に基づいて近似される。この正規化 soft-max 方策を用いると行動のランダム性が正規化され、探索的行動の行動価値関数のばらつきへの依存を弱めることになる。

正規化 soft-max 方策 (3.25) は、以下で定義される新しい逆温度を用いることによって、元の soft-max 方策 (3.24) と等価になる。

$$\beta(s) = \beta_0 \cdot \beta_l(s) = \frac{\beta_0}{\sqrt{E[Q(s,a)^2] - (E[Q(s,a)])^2}} \tag{3.26}$$

ここで、$\beta(s)$ は行動 a に依存しないことに注意する。行動価値関数のばらつきに対するランダム性の正規化により逆温度 β が状態 s によって決まるようになるため、$\beta_l(s)$ は逆温度の局所係数と呼ぶ。

行動価値関数 $Q(s,a)$ の行動に関する分散が小さい場合は、どの行動を取っても遷移する状態の価値があまり変わらないことが予想される。この場合、exploration は余り重要でないと考えられる。本研究でのメタ制御をでは、逆温度を大きくし

て行動選択のランダム性を抑え、一方、(3.24) 式の soft-max 方策 (逆温度が固定されている) によると、行動選択のランダム性がメタ制御を行う場合よりも大きくなる。逆に分散が大きい場合は、固定された逆温度を用いると高い確率で行動価値が最大の行動を選んでしまう。こうした重要な状態について、メタ制御では温度を高くすることで exploration を保持しているのである。

3.3.3 ランダム性の大域的制御

エージェントが環境の変化を認識したときに、exploration を行うことは重要である。逆に、環境が変化していないと信じている時には、現在の推定を重視して exploitation を行うべきである。つまり逆温度は環境変化の認識に基づいて制御されるべきであり、このような制御は以下によって実現できる。

$$\beta_g := \begin{cases} \alpha + (1-\alpha)\beta_g & (\text{if} \quad \hat{z}' = \hat{z}) \\ \beta_r & (\text{otherwise}) \end{cases} \tag{3.27}$$

ここで、$\hat{z}$ と $\hat{z}'$ はそれぞれ、行動前と行動後の不観測変数の推定値である。$0 < \alpha < 1$ は、β_g がその最小値 β_r から最大値 1.0 に近付く早さを決める定数である。実際に行動を取った後も不観測変数の推定が行動前と変わらない場合、エージェントは環境が変化していないと推測する。このような場合、β_g が徐々に増加しエージェントは exploitation をより重視するようになる。一方、エージェントが環境が変わったと認識した場合は、β_g が最小値になり、その結果、新しい環境に適応するために exploration を重視するようになる。

迷路探索問題のように環境が決定論的である場合、以下のような制御が有効に働く。

$$\beta_g := \begin{cases} \alpha + (1-\alpha)\beta_g & (\text{if} \quad z = \hat{z}) \\ \beta_r & (\text{otherwise}) \end{cases} \tag{3.28}$$

実際の不観測変数 z がその推定値 $\hat{z}$ と異なる場合、エージェントは環境が変化したと推測する。確率的環境では不観測変数の認識が推定値と異なることが、環境の確率的性質によって起こるため、この制御は適していない。

これらの制御方法によって、元の soft-max 方策における逆温度 β は以下のように置き換えられる。

$$\beta(s) = \beta_0 \cdot \beta_g \cdot \beta_l(s) \tag{3.29}$$

$\beta_l(s)$ が行動価値関数の分散を考慮して局所的なランダム性を制御するのに対し、β_g は環境変化の認識を用いて大域的にランダム性を制御する。β_g は逆温度の大域係数と呼ばれる。

3.3.4 探索ボーナス

本研究ではさらに、環境からの情報をより多く獲得できるように、exploration ボーナスを導入する。強化学習の目的は、状態と行動のペアに対して与えられる報酬の将来にわたる総和を最大化する方策を見つけることである。一般に、報酬関数はエージェントが行うべきタスクに応じて、エージェントの知識量とは無関係に決められる。一方で、報酬にエージェントの知識量に応じたボーナスを加えることによって、エージェントの探索行動を促進することができる。これを exploration ボーナスと呼ぶ。

exploration ボーナスのアイデアは、最初に Sutton [77] の Dyna システムで用いられた。Sutton の DYNA システムでは [77]、全ての状態行動対に対しその状態行動対を以前経験した時からの経過時間に基づいて、即時報酬に exploration ボーナスが加えられる。Moore と Atkeson [48] の exploration ボーナスは、よく知らない状態を高い価値を持つ架空の終了状態と関連させることで、エージェントの未知状態への訪問を促進する。また、Dayan と Sejnowski ら [18] の手法では、環境ダイナミクスを忘却する効果により、エージェントは価値関数の現在の推定値に関して最適でない行動を試すようになる。

本研究で用いた exploration ボーナスは、推定状態遷移確率のエントロピー

$$I(s,a) = -\sum_{s'} \hat{P}(s'|s,a) \log \hat{P}(s'|s,a) \tag{3.30}$$

に比例して、以下のように即時報酬に加えられる。

$$r^+(s,a) = r(s,a) + \varepsilon I(s,a) \tag{3.31a}$$

$$Q^+(s,a) = r^+(s,a) + \gamma \sum_{s'} \hat{P}(s'|s,a) V(s'(a)) \tag{3.31b}$$

エントロピーが小さい、すなわち状態 s において行動 a をとることによって得られる環境のダイナミクスに関する情報量が少ないと、その行動の選択確率を下方

修正する。すなわち、exploration の効果を弱める。一方で、行動によって環境のダイナミクスに関して多くの情報量が得られることが期待できる場合は、その行動の確率を上方修正する。すなわち、exploration を強く働かせる。エージェントはこの $Q^+(s,a)$ を用いて、行動選択を行なう。このボーナスは即時報酬に加えられるものであるが、最適方策および (3.8) 式のベルマン方程式には影響しない。したがって、価値関数の学習にはバイアスがかからない。

3.4. 強化学習アルゴリズム

まとめると、本研究で用いる手法は以下のようなアルゴリズムとなる。

1. エージェントをスタート位置に置く。
2. 決められたステップ数まで以下を実行。
   (a) 現在の観測状態 y に対する各不観測変数 $\hat{z}$ を、$k = \arg\max_i \bar{g}_i$ となるような $z_k = 1$ と推定する。ここで、$\bar{g}$ は (3.19) 式によって与えられる。
   (b) 現在の状態 y において可能な全ての行動 a について以下を実行。
      i. s と a の組について可能な観測状態 s' への遷移確率 $\hat{P}(y'|y,a)$ を (3.19) 式によって決める。
      ii. $s = [y, \hat{z}]$ と $\hat{P}(y'|y,a)$ を用いて、(3.21) 式によって $Q(s,a)$ を求める。
      iii. $s = [y, \hat{z}]$ と $\hat{P}(y'|y,a)$ を用いて、(3.31a) 式と (3.31b) 式によって $Q^+(s,a)$ を求める。
   (c) (3.8) 式に基づいて $V(s)$ を更新する[1]。
   (d) (3.29) 式によって $\beta(s)$ を求める。
   (e) $Q(s,a)$ を $Q^+(s,a)$ に、β を $\beta(s)$ に置き換えて、(3.25) 式に従って $P^\pi(a|s)$ を計算する。

[1]確率的環境では、徐々に近付けるような更新が望ましい。

(f) エージェントは $P^{\pi}(a|s)$ の確率にしたがって行動 a を実行する。行動の結果、新しい状態を y'' とする。

(g) (3.18) 式を用いて y'' に対する十分統計量を更新する。

(h) (3.20) 式によって s と a の組に対する実効的データ数をインクリメントする。

(i) y'' がゴールでないならば $y \leftarrow y''$ として (a) へ。ゴールならば終了。

3. (3.20) 式によって全ての状態行動対に対する実行的データ数を減少させる。

上記は 1 エピソードに関するアルゴリズムである。学習エピソードの列を開始する前に、$V(s), \langle z \rangle_D, \tau$ などを初期化する必要がある。

3.5. シミュレーション実験：迷路探索問題

3.5.1 問題設定

上で述べた手法を 2 次元迷路探索問題に適用し、シミュレーション実験を行なう。ここで用いた迷路は、Sutton [77] の論文で発案された迷路を元に Dayan ら [18] が考案した迷路の一部を変更したものである。エージェントは、16×16

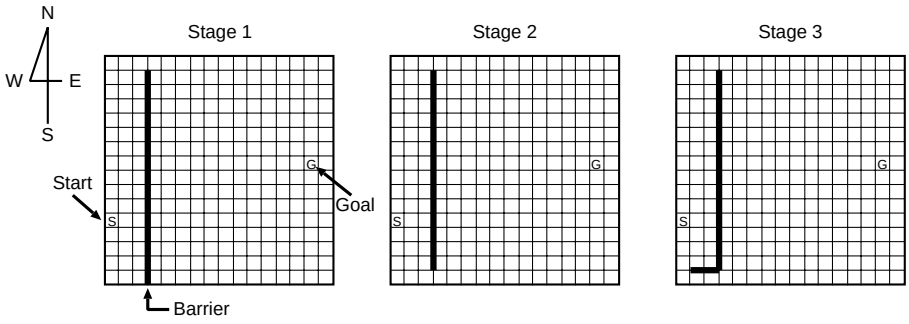

Figure 3.2 迷路探索問題

グリッドの 2 次元迷路内をスタート位置からゴールまで移動する。各状態におい

て、エージェントは $a \in \{N, S, E, W\}$ の 4 つの行動を取ることができる。各行動に対して -1 の即時報酬、すなわち 1 のコストが与えられる。各エピソードは固定されたスタート位置から始められ、エージェントがゴールに到達するか、あるいは最大ステップ数の 200 ステップ進むと終了する。エージェントは、ゴールまでのコストを最小化する、つまりゴールへの最短経路を通るように学習する。行動可能な範囲は迷路の中だけで、迷路の外に出る行動は取れない。

迷路には両方向性のバリアが存在し、エージェントはバリアを越えることはできない。バリアに向かって進もうとすると、-1 の即時報酬は与えられるが状態は変化しない。バリアは直接観測できない、すなわち、バリアの有無は行動を取った時に状態が変化するかどうかによってのみ認識可能である。バリアの存在は決定論的に決まるが時間とともに変化するため、確率的事象と仮定される。バリアの存在は不観測確率変数として表される。エージェントがある状態である行動を起こした時に起こる事象は、次の状態に移動するかどうか、つまりバリアが存在するかどうかという 2 項の事象であり、各不観測変数に対して取り得る値は 2 になる。したがって、エージェントは全ての不観測変数 (あるいは状態遷移確率) に対する二項確率モデルをベイズ推定によって近似する。これが、環境モデルの推定である。

本研究では、環境変動に対するエージェントの行動適応性を調べるために、試行中にバリアの位置を変える。この変化を許容することが、バリアの有無が確率的であるとする仮定に相当する。試行 (学習エピソードの列) を 3 つのステージに分ける。それぞれのステージでの迷路の状態を図 3.2 に示す。最初のステージ (1 $\sim$ 400 学習エピソード) では南北方向のバリアがあり、一番北側だけが通れるようになっている。この段階での最短経路はバリアの北側を回り込んでゴールに向かう経路で、最小ステップ数は 32 である。ステージ 2(401 $\sim$ 800 学習エピソード) に入ると、バリアの一番南側が取り除かれる。これによって最短経路はバリアの南側を通るパスに代わり、この時の最小ステップ数は 26 となる。ステージ 3(801 $\sim$ 1200 学習エピソード) では、スタート位置から下側のゲートへの経路を阻害するような新しいバリアが出現する。ここでは最小ステップ数は変わらないが、最短経路の範囲が狭まり、スタート位置からまっすぐ南に進む経路のみが最短経路となる。

本研究で用いたモデル同定強化学習法では、バリアの存在を忘却する効果がある。エージェントは前回試した後でしばらく時間が経つとバリアがなくなっているかも知れないと期待するため、同じ最短経路でもバリアに沿って移動する経路を取りやすい。したがって、たとえゴールまでの距離が同じであっても、バリアに沿ったグリッド点の方がバリアから離れた点よりも高い価値関数を持つようになる。実際にステージ 2 で獲得される経路は、まずバリアに向かい、その後でバリアに沿って南下し、バリアを回り込むものである。したがって、その後にステージ 3 のようになると、さらに回り込むことを試みるため、スタート位置からまっすぐ南下する最短経路を得るのは容易ではない。

状態価値関数 $V(s)$ は 0 で初期化する。価値関数の推定初期値は初期の行動価値推定値に関してバイアスを持つため、exploration を促す簡単な手段として用いることができる。価値関数の初期値を実際に与えられる報酬よりも大きい値に設定すると、いずれの行動が最初に選択されても、受け取る報酬は最初の推定量よりは小さい。したがってエージェントは報酬に「失望」して、別な行動へと切替える。その結果、価値関数の学習が収束する前にすべての行動が数回試みられる。この楽観的初期値 [93] は、exploration を促進するための簡単な方法の一つであるためここでも用いている。しかし、楽観的初期値は一時的にしか exploration を促進しないため、本研究のタスクのように動的な環境の場合には、この方法による exploration では不十分と考えられる。

3.5.2 実験結果

上で述べた迷路探索問題に、逆温度メタパラメータの制御を行うモデル同定強化学習法を適用する。図 3.3に学習時のエージェントの行動ステップ数を示す。図 3.3の横軸は学習エピソード回数で、縦軸は 10 エピソード毎の平均ステップ数 (上図) と各エピソードにおけるステップ数 (下図) である。ステージ 1 ではステップ数のバラつきが大きいことから、行動のランダム性が高いことが分かる。これは、学習初期は状態価値関数が真の値と大きく異なっているため、行動がランダム (でたらめ) であることによる。楽観的初期値の効果もある。また、後のステージよりも経路中にバリアに沿って進む距離が長いため、バリアにぶつかる回数が多くなっていることも影響している。一方で、ステージ 2 とステージ 3 では状態価値

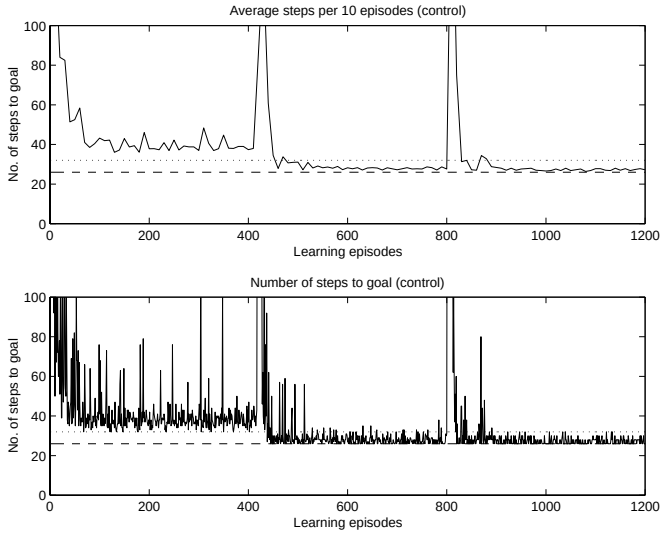

Figure 3.3 逆温度をコントロールした場合の行動ステップ数

関数が真値に近付き、その値に基づいてより適切な行動選択が行われるため、ステップ数のバラつきが小さくなる。また、バリアに沿って移動する距離が短いことも影響している。図 3.3の上図から、平均ステップ数がステージ 2 で減少しており、エージェントが環境の変動に適応し、新しい最短経路を発見できていることが分かる。ステージ 3 で新しいバリアが作られると一時的にステップ数が増えてしまう。これは先に述べたように、エージェントはバリア沿いに進む経路を取りやすいが、その経路が阻まれたことを反映している。しかし、学習が進行するとバリアに沿わずに進む経路を獲得し、平均ステップ数はステージ 2 の時と同じに戻る。すなわち新しい最短経路を獲得する。図中の点線と破線はそれぞれ、ステージ 1 での最短経路 32 ステップとステージ 2 および 3 での最短経路図 26 ステップを示す。下図から、エージェントが全てのステージで最短経路を獲得できていることが分かる。

本手法における逆温度制御の意義を調べるために、逆温度パラメータに定数を用いた場合と比較する。図 3.4と図 3.5は、それぞれ逆温度を大きい値 $(\beta = 100)$ と小さい値 $(\beta = 1.0)$ に固定して、同じ迷路探索問題に適用した結果である。逆温度を固定した場合においても、exploration ボーナスは用いている。用いないと、逆温度が大きいエージェントは環境変化にほとんど適応できない。逆温度として

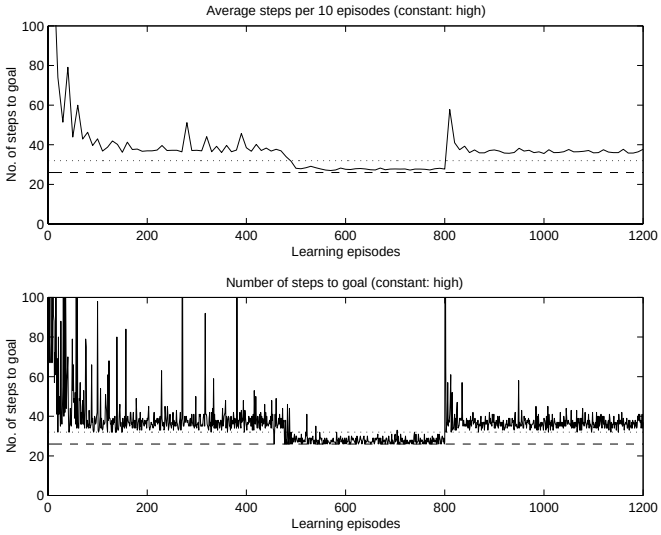

Figure 3.4 逆温度を大きい値に固定した場合の行動ステップ数

大きい定数を用いた場合、行動のランダム性が低くなるため、exploitation が重視され、どのステージにおいても短い経路でゴールすることができる。しかし、図 3.4の上図を見ると、ステージ 3 での平均ステップ数がステージ 2 の時よりも大きくなっている。このことから、エージェントはステージ 3 での最短経路を発見できておらず、バリアの北側を通る準最適経路を通っていることが分かる。これは、逆温度が常に大きく設定されているため十分な exploration が行われておらず、環境の変動に追随できないためである。一方、逆温度として小さい定数を用いると、新しい経路を見つけることはできるものの、行動のランダム性が高いために、平均ステップ数は大きくなる (図 3.5)。このように、逆温度を大きく固定することは、exploitation を重視することであり、逆に小さく固定することは exploration を重視することである。いずれにおいても exploitation と exploration の両者を同時に実現することができない。

図 3.6に 3 エージェントの平均パフォーマンスを示す。図の縦軸は、異なる 100 個の初期状態から学習を行った時の平均行動ステップ、横軸は学習エピソード数である。

次に、タスク遂行中のエージェントが、実際にどのような経路を通っているか

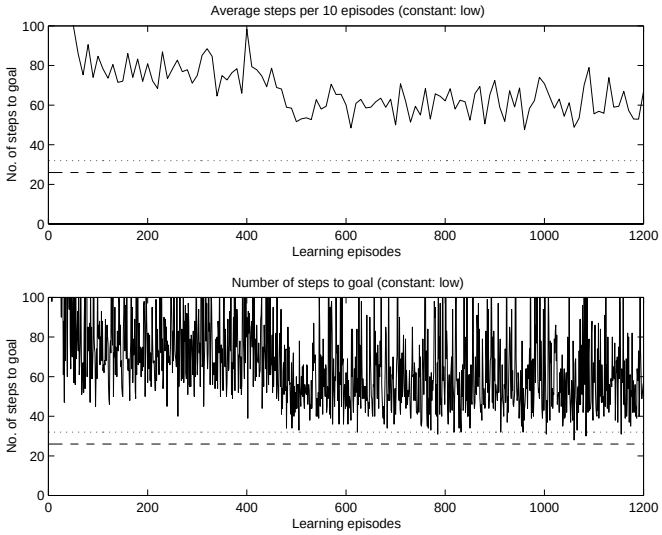

Figure 3.5 逆温度を小さい値に固定した場合の行動ステップ数

を調べた。図 3.7は、各状態についてエージェントが通った回数を累計し、その対数を示したものである。図の左から右へ順にステージ 1,2,3 での結果を示した。また、明るい色の状態程通った回数が多い。ここでは、環境をある程度学習した後のエージェントの行動をみるため、各ステージの後半の 100 学習エピソード中のエージェントの行動を示している。図 3.7は上から逆温度を制御した場合、逆温度に大きい定数を用いた場合、小さい定数を用いた場合の結果である。温度を制御した場合、すべてのステージで最短経路を発見できていることが分かる。つまり、環境の時間的な変動にうまく適応できていると言える。また、特にステージ 3 において、バリアを越えるまではほぼ確実に同じ経路を取るのに対し、バリアを越えてからゴールまでの間は環境を探索し、様々な経路を通っていることが分かる。このことから、逆温度を状態に応じて変化させることによって、環境の空間的な変化にも適応できていると言える。逆温度に大きい定数を用い、温度を低く設定した場合、殆んどいつも同じ経路を辿っている。特に、ステージ 2 でバリアの東側の探索が不十分であるため、ステージ 3 での最短経路を見つけることができず、以前良かったポリシーを用いてバリアの北側を回る経路を通るようになる。逆に、逆温度を小さくすると、最短経路を選択的に用いることはなく、ゴー

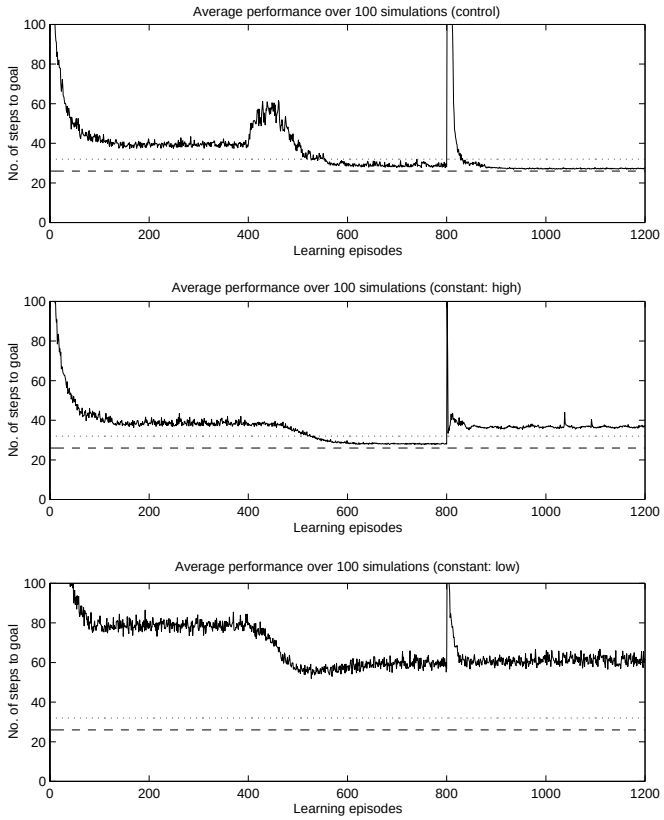

Figure 3.6 各エージェントの平均行動ステップ数

ルへの可能な経路を万遍なく通るようになる。実際に、ステージ 3 ではバリアの北側と南側を回る 2 つの経路をほぼランダムに選んでいることが分かる。

図 3.8は、1 回の学習中の逆温度の大域係数 β_g の変化を示したものである。エージェントが環境の変化を認識すると、β_g を小さくして新しい環境に適応するための exploration を促す。一方、環境変化が認識されないときは、β_g を増加させて exploitation の割合を強める。図 3.9に、各グリッド点での逆温度局所係数の逆数、つまり $1/\beta_l(s)$ を示す。バリアから離れたグリッド点では、逆温度が大きいためエージェントは exploitation をより好むようになる。一方、バリアと隣り合ったグリッド点では逆温度が小さくなるため、エージェントは exploration をより好むようになる。つまり、バリアがなくなることを期待するようになる。

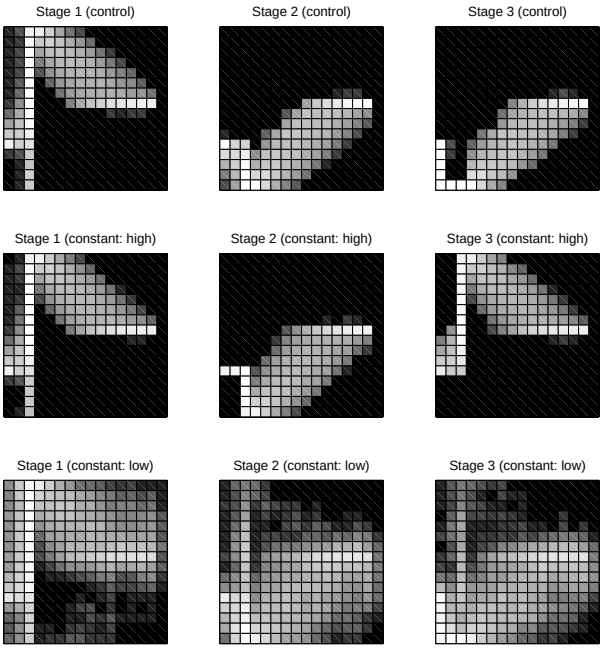

Figure 3.7 学習後のエージェントの位置の分布

3.6. まとめ

本章では、部分観測環境におけるモデル同定強化学習法について議論し、環境のダイナミクス、すなわち状態遷移確率の近似にベイズ推定を用いる手法を提案した。本手法では、無情報事前知識を用い、過去の経験に基づく推定の忘却効果を導入した。この忘却により、環境の探索が促進される。また、exploration と exploitation の制御について議論した。exploration と exploitation のバランスを制御するために、エージェントの行動選択に逆温度メタパラメータの制御機構を導入した。この制御は、行動価値関数のばらつきと環境変化の認識の両方に基づいて行われた。また、探索を直接的に促す手法として、エントロピーに比例した exploration ボーナスを導入した。

本手法をバリアのある 2 次元迷路探索問題に適用し、逆温度の制御を行わない手法と比較した。実験により、逆温度を状態に応じて制御することによって、環境の変動にうまく適応できることが分かった。

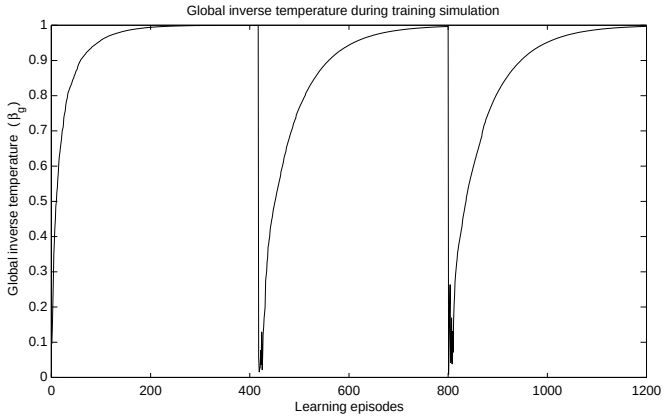

Figure 3.8 逆温度大域係数 β_g の時間変化

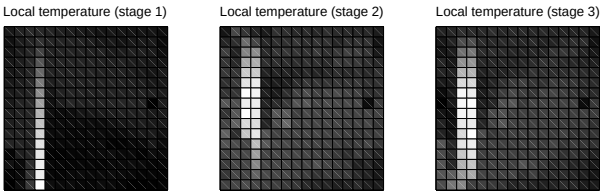

Figure 3.9 逆温度局所係数 $1/\beta_l(s)$ の時間変化

Chapter 4

モデル同定強化学習の脳内モデル

本章では、第3章で議論したモデル同定強化学習を実現する脳内モデルを提案し、認知心理学、神経生理学による知見に基づき議論する。

近年コンピューターやロボットの学習にも適用されている強化学習理論は、心理学の分野から生まれたもので、本来ヒトをはじめとする動物についての学習理論であった。池のそばで手を叩くと、泳いでいる鯉が寄ってくる。池の鯉は、音がしたときに人影のそばに行くと餌をもらったという経験から、こうした行動を学習する。動物は、ある行動で良い結果が得られればその行動を取り続けるし、失敗すれば別の行動を試みる。自発的な行動はその行動の引き起こす結果によって変容する。その過程を明らかにしたのが 1911 年のソーンダイクによる道具的条件付けの研究である。彼は、仕掛けのある箱の中に空腹のネコを入れ、ネコが仕掛けを操作して外に出る過程を観察し、ネコにとって満足のいく結果を生じる行動の生起する頻度が学習によって上昇するという「効果の法則」を導いた。これを基に、統制された条件下における動物行動の学習であるオペラント条件付け型学習が提案された。

試行錯誤によって偶発的に報酬を得る行動を見つけると、その行動を生起する頻度が上昇する。この学習過程、特に報酬が遅れて与えられる状況 (遅延報酬課題) における行動変容を数理的に説明ようとするのが強化学習理論である。最近の神経生理学研究によって、中脳ドーパミン系の活動が強化学習における報酬予測誤差と類似の働きを持つことが示された [69, 87]。また、大脳基底核が強化学習法の 1 つである Actor-critic 法を実現する構造を持つと提案する研究もある [6]。

さらに、最近の脳活動計測装置を用いたいくつかの研究は、脳におけるいくつかの局所領域の機能と強化学習で用いられる関数が対応づけられることを示唆している。つまり、行動心理学研究に由来する強化学習は、数理科学や工学分野で学習理論モデルとして発展し、神経生理学、認知心理学の研究によってその脳内実現性が示唆されている。本章では、第 3章で述べたモデル同定強化学習法を対象とし、現在までの神経生理学、認知心理学の知見に基づいた脳内モデルを提案する。第 3章で述べた強化学習法では、未知の環境を同定し、それを用いて報酬の予測を行い、行動を選択する必要がある。また、環境の変化を認識し、行動選択の戦略を切替える必要がある。さらに、環境に存在する隠れ変数を観測に基づき推定する必要がある。本モデルでは、環境モデルの同定とその評価に背外側前頭前野 (dorsorateral prefrontal cortex: DLPF) が、行動選択の制御に前部帯状回 (anterior cingulate cortex: ACC) が、隠れ変数の推定に前部前頭前野 (anterior prefrontal cortex: APF) が関わると想定している。

4.1. 行動選択制御の脳内モデル

4.1.1 行動のランダム性と青斑核

強化学習法でしばしば用いられる soft-max 方策によると、行動のランダム性は逆温度メタパラメータに依存している。第 3章では、逆温度という 1 つのメタパラメータで exploitation と exploration のバランスを決定し、かつ逆温度の値を環境の推定によって制御する手法を提案した。強化学習の学習様式を特徴付けるメタパラメータは従来手で決められており、制御方法も例えばアニーリングのように天下り的に与えられていた。メタパラメータの自動制御に関する研究は工学的に重要であるが、第 3章の研究は以下で述べるように、神経生理学の知見に動機付けられたものでもある。Usher ら [85] は、ノルアドレナリンを伝達物質とする青斑核 (locus coeruleus: LC) ニューロンの発火パターンが、行動パフォーマンスと関わることを観察した。行動正答率が高いときはターゲット刺激に対する選択的な一過性発火が見られる。一方、妨害刺激に反応してしまいパフォーマンスが悪くなるときには、高いレベルでの持続性発火が続く。正しい行動を取る時は、ターゲットに対する選択的注意レベルが高くなっていると考えられる。彼らは、

LC ニューロン間の電気的結合の強さを変えることによって発火パターンが変わり、一過性発火が exploitation に、持続的発火が exploration を引き起こすという仮説を提案した。このアイデアは、第 3章で提案した、exploitation-exploration バランスを 1 つのパラメータ、すなわち逆温度によって制御する手法と類似しており、逆温度が LC で表現されていると考えられる。

4.1.2 戦略の切替えと前部帯状回

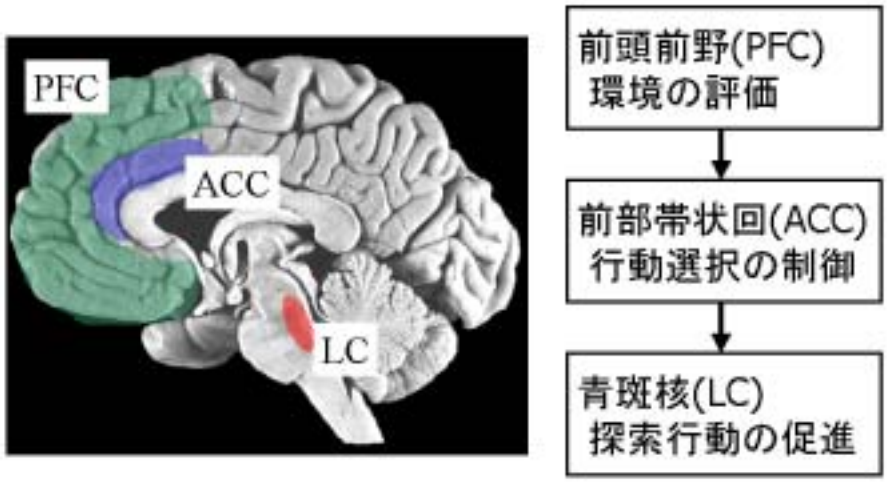

Figure 4.1 行動選択の脳内モデル

最近の生理学研究 [55] により、LC が前部帯状回 (ACC) から投射を受けていることが明らかになった。このため、ACC が LC ニューロン間の電気的結合を制御することにより、その活動度を調整している可能性が考えられる。ACC (ブロードマンの 24,32 野) は前頭葉正中表面に位置し、脳梁の上部を覆う領域である。ACC は扁桃体からの密接な投射を受け、自律機能や運動出力に関与する領域に強い繊維投射があるため、情動に基づく行動の実行システムとして機能していると考えられてきた。また、最近の神経解剖学的研究によって、ACC が複数の機能を持つ部分領域を持つことが示唆されている。本モデルでは、LC ニュー

ロンが逆温度を表現し、ACC が逆温度の局所係数と大域係数によって逆温度の制御、つまり行動選択に関わると仮定する。

最近の単一ニューロンの電位記録実験 [72] は、サルの帯状運動野 (cingulate motor area :CMA [1]) ニューロンが報酬依存の意思決定に関係することを示唆している。得られる報酬が一定である間、サルは複数の可能な行動のうち報酬が貰える特定の行動を取り続けるが、報酬量が減少すると異なる行動を試すようになる。つまり、能動的に行動ルールの切替えを行い、この時に CMA 吻側部のニューロンが活動することが明らかになった。さらに、この領域に GABA 作動性ニューロンの促進剤であるムシモールを注入して機能阻害すると、スムーズな切替えが出来なかったり必要もないのに切替えてしまう行動異常が見られた。これらの結果は、CMA が期待報酬量と実際の報酬量の差分に依存した随意的な行動選択に重要な役割を果たしていると考えられる。同様のタスクを用いたイメージング研究 [11] でも、被験者が能動的に行動ルールを切替えるときに ACC の活動が観察された。

ACC はまた、エラーや行動のコンフリクトを検知したときに活動することが知られている。電気生理学研究の分野では、ACC で誤り関連活動が起こることが古くから知られている。最近のイメージング研究 [10] によると、主に ACC の吻側後部領域が誤り検知に関連して活動する。これらの結果は、ACC が自身の行動の結果、すなわち誤りによって環境の変化を検知し、それによって行動ルールを変化させていると予想させる。そして、この行動選択の制御方法は、第 3章で述べた学習スキーマにおける逆温度の大域係数を用いた制御手法と一致する。一方、ACC は誤りそのものではなく矛盾した行動のコンフリクトの検知によって活動するという説もある [8, 12]。これを、コンフリクトモニタリング説という。視覚弁別タスクを用いたイメージング研究で、誤った行動を取った時よりも、むしろ行動のコンフリクトが強い状況で正しい行動を取っている時に、より強い一過性の活動増加を見せることが明らかになった。ACC はまた、新奇な環境において活動する。系列運動学習タスクを用いた PET (positron emision tomograpy) 研究 [36] によると、ACC は学習後の自動的な行動実行時には活動しないが、新しい

[1]サルの CMA は大脳皮質内側部の帯状溝を覆う領域で、ヒトの ACC 領域と重なる部位である。

系列を学習している間には活動することが分かった。この結果から、ACCが行動(反応)の準備に注意が必要な時に活動すると考えられていた。しかし、Jueptnerら [37] は、学習後の系列タスク遂行時に次の行動を予測し注意を向けた場合には、ACCが活動しないことを示した。これらの結果から、被験者の状態、つまり行動のコンフリクトレベルや環境に対する知識量に依存してACCの活動度が変わるといえる。第3章で述べた学習スキーマでは、エージェントの状態に依存した逆温度の局所係数 $\beta_l(s)$ を用いて、行動のランダム性が制御される。つまり、現在の方策に関する行動価値関数のばらつきが大きいと、局所係数が大きくなり探索的行動が促される。行動価値関数のばらつきの大きさは、主に方策のばらつきの大きさによって決まる。現在の方策がばらついている時に探索を促進することとなり、このような状況は、よく知らない、あるいはコンフリクトの強い状態といえる。したがって、上で述べたようなACCの活動は、逆温度局所係数による制御手法と一致すると考えられる。

4.2. 環境の評価と前頭前野

行動価値関数のばらつきを求めるためには、環境の評価値が必要になる。ACCは主に前頭前野 (prefrontal cortex: PFC) からの投射を受けるため、環境の評価がPFCで行われていると想定する。

脳の皮質連合野はその機能と構造から、外界からの種々の感覚入力を分析し且つ統合処理をして判断する領野である後連合野 (頭頂・側頭・後頭) と、その判断に基づいて外界に対して能動的に働きかける役割を持つ前連合野 (前頭前野: PFC) に分けることができる。連合野は動物が高等になるにつれて領域的に広くなり、特にPFCで顕著である。後連合野が感覚入力を直接受けとるのに対し、PFCは他の連合野で処理された感覚入力を受けとる。PFCの主な出力先は、線条体や運動連合皮質のような運動系である。PFCは、意思決定、行動抑制、行動のプランニング、行動の評価、作業記憶の保持など、様々な高次機能を管理していると考えられている。PFCの機能は、単一の理論では説明できないため、機能別にいくつかの部分領域に分割して検討するべきである。ここでは特に、背外側前頭前野 (DLPF) と前部前頭前野 (APF) の2つの領域に注目し、これらの領域の機能と第

3章で述べたモデル同定強化学習における主な関数との関係について議論する。

4.2.1 環境モデルと背外側前頭前野

DLPF の研究は、主に作業記憶、つまり必要な情報のアクティブな保持機能に関して行われてきた。Rao ら [56] は、視覚誘導サッケード (視機性眼球運動) タスクを行っている際のサルの DLPF ニューロン活動を記録した。このタスクは、まずはじめにサンプル刺激が注視点に表示され、遅延時間 (what delay) 後に、サンプル刺激と妨害刺激が 4 つの提示位置のうちの異なる位置にそれぞれ提示される。さらに遅延時間 (where delay) が経つと、サルはサンプル刺激が提示された位置にサッケードすることによって報酬が貰える。つまり、what delay は目的指向行動のための部分的な情報を、where delay は適切な行動そのものを保持している時間である。タスク遂行中、DLPF ニューロンは、what と where のどちらかあるいは両方の遅延時間の間、持続的な発火を示した。つまり、DLPF には状態依存と行動依存の持続的活動を示すニューロンが存在し [31]、これは DLPF ニューロンが状態/行動依存の何らかの情報処理に関わることを示唆する。

また、最近の研究は、DLPF ニューロンが将来与えられる報酬の質や量を予想していると指摘している [88, 43]。遅延反応タスク遂行中のサル DLPF ニューロンは、好ましい報酬が予想できるときに、好きでない報酬の場合よりもより強く活動する [88]。記憶誘導眼球運動タスクを用いたその後の研究は、DLPF ニューロンが報酬の量が多いときに強い発火を見せることを示した [43]。これらの実験結果は、DLPF が状態あるいは行動に依存して活動し、その活動度が累積報酬和の推定値を表現していることを示唆しており、これは、強化学習における状態価値関数や行動価値関数と考えることができる。

最近の見解では、DLPF が目的指向型行動を達成するために必要な情報を、オートマトン、すなわち状態遷移を表現したカスケードネットワークとして維持していると考えられている [80]。サルを用いた生理実験によって、DLPF の運動関連神経活動が記録された [31]。さらに、最近の脳活動計測研究 [53] は、DLPF が情報の保持よりもむしろ、作業記憶によって保持された情報を用いた行動系列の準備に関わると指摘している。行動をプランニングするためには、自身の行動によって環境がどのように変化するかを予測する必要がある。したがって、モデル

同定強化学習における環境モデルが DLPF で表現されていると考えられる。

4.2.2 隠れ状態の推定と前部前頭前野

APF は PFC の中でも最も前方に位置し、他の霊長類に比べてヒトでの発達が最も顕著な部位である。APF は破壊実験が困難であるため古くからのデータが少ないが、最近の非侵襲脳計測実験によってその機能が明らかになってきている。Koechlin らは、サブタスク遂行中に主タスクの作業記憶保持が必要なブランチングタスクを用いて、タスク遂行時の脳活動を計測した [42]。タスク条件として、主タスクとサブタスクが周期的に提示される予測可能条件と、2 つのタスクがランダムに提示されるため予測が不可能なランダム条件の 2 種類が設定し、APF がランダム条件で活動することを示した。また、明示的分類タスクを用いた他のイメージング研究 [75] は、行動ルールの変更に関連して APF がすることを明らかにした。これらの結果は、明示的な手掛かりを用いずに能動的に行動ルールを切替える際に APF が活動すると提案している。しかし、このような切替えは環境変化を推定することによって引き起こされるため、APF が不観測状態の推定に関わると解釈することもできる。

4.3. 脳内仮説モデルのまとめ

本章では、脳機能部位と第 3章で述べたモデル同定強化学習における様々な関数を対応付けることにより、モデル同定強化学習の脳内仮説モデルを提案した。

第 3章の手法は、青斑核 (LC) ニューロンの活動が強化学習の逆温度メタパラメータと類似の役割を持つという神経生理学研究に動機付けられたものである。すなわち、逆温度 β が LC で表現されていると考える。この逆温度は状態価値の分散に基づく局所係数 β_l と、環境変化の認識に基づく大域係数 β_g によって制御される。LC ニューロンの活動パターンが前部帯状回 (ACC) からの投射によって制御されているという可能性から、これらの係数は ACC で表現されるものとする。行動選択の不確実性と環境の変化によって行動選択の戦略を切替える仕組みは、この LC と ACC のネットワークによって実現される。

行動の選択は、状態の良さを評価することによって行われる。また、環境の変

化は、環境モデルから予測される結果と実際に環境から与えられる結果の差分として認識される。つまり、行動戦略を適切に切替えるためには、環境の良さと状態遷移を学習、表現すること (環境モデル) が必要となる。実際、第 3章で議論した強化学習モデルでは、逆温度の制御係数の計算に、行動価値関数 $Q(s,a)$ と状態遷移確率 $P(s'|s,a)$[2]を用いている。本章で述べた脳内モデルでは、環境モデルの同定とその評価を背外側前頭前野 (DLPF) が行う。DLPF で表現されるこれらの情報は、ACC への直接結合によって伝達され、逆温度制御係数の計算に用いられる。さらに、環境に観測できない隠れ変数 z が存在する場合、環境モデルには隠れ変数の推定値 $\hat{z}$ が必要となる。隠れ変数を観測に基づいて推定する部位としては、前部前頭前野 (APF) を想定した。

[2]迷路の経路探索問題の場合、隠れ変数の期待値 $\hat{z}$ と等価になる。

Chapter 5

強化学習の脳内モデルの検証実験

本章では、第 4章で提案したモデル同定強化学習の脳内モデルを検証するための認知心理学研究について述べる。本研究では、非侵襲脳計測機器である核磁気共鳴画像 (functional magnetic resonance imaging: fMRI) を用い、強化学習タスク遂行時のヒトの脳活動を計測し、その解析を行った。第 4章で議論したモデルでは、神経修飾物質の役割や、複数の脳機能部位の情報伝達経路についても言及している。しかし本研究で用いた fMRI は、酸化ヘモグロビンの増加から血流量の増加した部位を測定するという間接的な測定法であり、賦活部位の時間変化や事象関連活動を調べることは困難である。また、複数の被験者の画像データを統合して解析するために、各被験者からのデータを同じテンプレートを用いて標準化し、統計処理を行う。そのため、大脳基底核のような複雑で個人差のある構成を持つ領域の活動部位を特定することは難しい。これらの理由から、本章では特に、大脳皮質の広い領域で定常的に計測される活動に焦点を当てて報告する。

5.1. 核磁気共鳴画像実験の概要

5.1.1 被験者

本実験には、16 名 (男性 13 名、女性 3 名、22-27 歳) の被験者が参加した。実験前に、すべての被験者から文書によるインフォームド・コンセントを得た。被験者はすべて大学院生で、神経系あるいは精神系の障害を持つ者はいなかった。

被験者には、実験タスクのスコアに応じた謝金が支払われた。被験者には、タスクのスコアに応じて謝金が支払われることが実験前に示されるため、本タスクは報酬依存であるといえる。

5.1.2 タスク設定

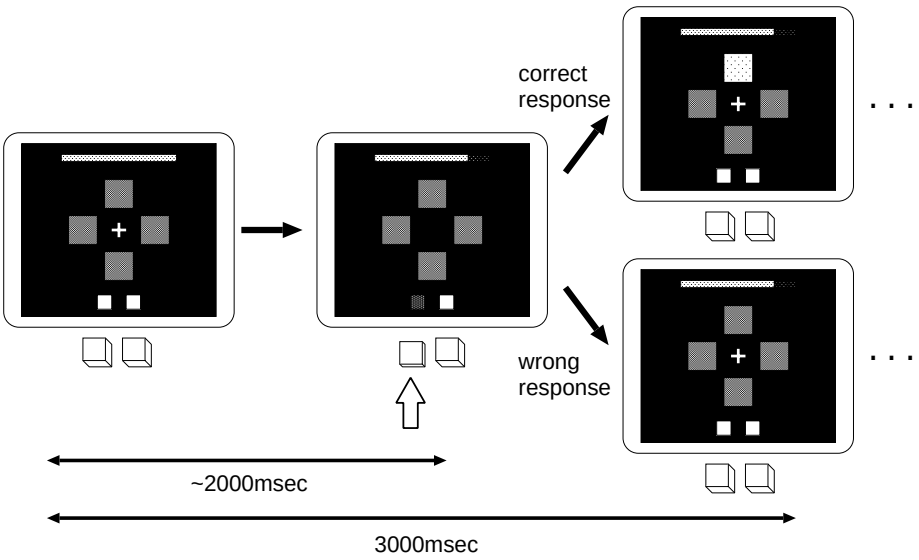

Figure 5.1 視覚刺激と反応ボタンによる系列学習タスク

本研究では、ボタン押しによる系列学習タスクを用いる。被験者は、スクリーンに表示される視覚刺激に対して 2 つの応答ボタンのどちらかを押し、刺激応答間の関係を学習する。図 5.1に本実験で用いたタスクを示す。初めに、スクリーンの中心に固視点と、その上下左右方向に 4 つの灰色の四角形が提示される。また、スクリーンの上部には各タスクにおける残りの試行回数を示す緑色のトライアルバーが、下部には 2 つの応答ボタンに対応した 2 つの灰色の反応確認ボタンが提示される。各試行において刺激は 3000 ミリ秒毎に提示されるが、固視点は始めの 2000 ミリ秒間だけ提示される。被験者の応答は、効き手の手元に用意された左右のボタンのうちいずれかを押すことであり、固視点が表示されている間に行動を行う。つまり、固視点は視線を固定する役割を持つのと同時に、行動

を行う手掛かりの役割も持つ。被験者がボタン押し行動を行うと、対応する反応確認ボタンが緑色に変化し、トライアルバーの長さが 1 試行分短くなり、固視点が消滅する。その直後に、固視点の周りを囲む 4 つの四角形によって次の状態が表示される。被験者が正しいボタンを押すと、1 つの四角形の色が変わるが、間違ったボタンを押した場合は何も変化しない。被験者が与えられた時間 (2000 ミリ秒) 内にボタン押しを行わないと、間違った行動と取ったと見なされる。状態は図 5.2 に示すような 4 つの四角形の色のパターンとして表現され、上部の四角形から時計周りに色が変わっていく。1 周目は灰色から赤色へ、2 周目は赤色から灰色へと変化し、初期状態に戻った状態をゴールとみなす。どちらのボタンが正しいかは、状態ごとに異なる。正しいボタンはあらかじめ決められているが、被験者には教示されない。つまり、被験者はゴールに到達するために、自身の行動が正しかったか間違っていたかを示す試行毎のフィードバックを用いて、8 つの行動系列を学習する。このタスクは、終端状態に到達するような状態遷移を試行錯誤によって学習する、オートマトンの学習とみなすことができる。

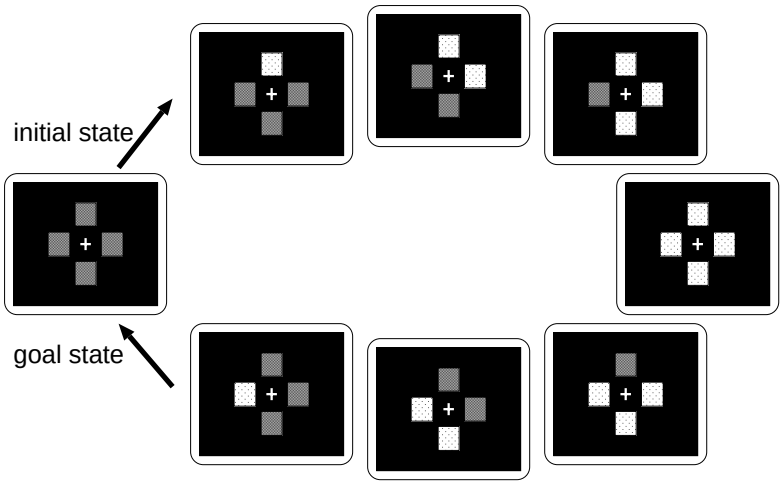

Figure 5.2 刺激の状態遷移図

本実験では、異なるオートマトンを学習対象とした 2 種類のタスク条件を用いた。メモリー (MEM) タスクでは、状態遷移が各状態に対して決定論的であり、被験者はある決まった 8 つの応答系列を記憶することによってタスクを行う。一

方、マルコフ決定過程 (MDP) タスクにおいては、状態遷移は一階のマルコフ過程である。つまり、各状態で正しい行動を取った場合でも、状態が変化する (成功する) 確率は 85%、その状態に留まる (失敗する) 確率は 15% に設定する。この状態遷移確率は時間によって変化しないため、被験者が状態遷移を確信していれば、その行動選択は MEM タスクのときと同じようにほとんど自動的なものとなる。各被験者は、3 セッションの MDP タスクと 1 セッションの MEM タスクからなるタスクを 2 回行った。1 セッションは 20 試行で約 6 分である。各セッションの間には、6 試行ずつのコントロールタスクを行う。コントロールタスクは、左右どちらかに提示される赤色の四角形を認識し、対応する側のボタンを押すという、記憶や学習を必要としない単純な反応タスクである。MDP タスクは 3 セッションに分割して行うが、その間に状態遷移やその確率が変化することはない。そのため、被験者はコントロールタスクの間も、学習した系列の記憶を保持しておく必要がある。各セッションの前には、タスク条件とセッションの番号が視覚メッセージとして提示される。また、セッション後には、各タスクの結果としてゴールへの到達回数が示される。

被験者は、スキャンの前にタスク内容に関して文書と口頭で説明を受け、タスクの練習を行なった。練習に用いたタスクは実際の実験タスクと同じ長さであるが、MDP タスクの状態遷移確率には実験タスクよりも簡単なものを用いた。

タスクは Presentation 0.50 (NeuroBehavioral Systems, Inc., San Francisco, CA) というソフトウェアによって作成し、刺激はコンピュータディスプレイを鏡で投射することによって被験者に示した。

5.1.3 撮像手続き

実験には、1.5 テスラの全身スキャナ (Magnetic Eclipse; Shimadzu Marconi, Kyoto, Japan) を使用した。機能画像は、酸素化ヘモグロビンレベル依存 (BOLD) 強度を持つ T2*強調のエコープラナー画像 (EPIs) として撮像した (TE, 55msec; flip angle (FA), 90°)。各ボリュームは、3.0 秒毎に連続して獲得し、5mm 厚の 28 スライスを 1 ボリュームとした (imaging matrix size, 64×64; field of view, 192×192 mm)。刺激は、スキャンと同期して提示された。最初の 4 枚の EPI 画像は、T1 平衡の影響を避けるために取り除いた。また、解剖学的位置測定に用

いるため、各スキャンの最初に高解像度の T1 強調 3 次元ボリュームを撮像した (voxel size, $1 \times 1 \times 1\text{mm}^3$)。

5.1.4 fMRI データの解析手法

fMRI 画像データは、statistical parametric mapping 99 (SPM99, Wellcome Department of Cognitive Neurology, London, UK) を用いて解析した。解析プログラムは Matlab 6.1 (Mathworks Inc., Sherborn, MA) で実行した。画像データの前処理として、初めに被験者毎の EPI 画像を各被験者の最初の画像を参照画像として剛体変換を用いて再調整し、体の動きによる画像のぶれを補正する。次に、coregistration を用いて、EPI 画像と T1 強調解剖画像の位置を揃える。これによって、異なるモダリティの画像の位置を合わせることができる。この解剖画像を MNI (Montreal Neurological Institute) 標準脳をテンプレートとして標準化し、このパラメータを用いて各 EPI 画像の標準化を行う。この標準化は、複数の被験者の形や大きさの異なる脳からのデータを統合して解析することを可能にする。標準化された EPI 画像は $2 \times 2 \times 2$ mm^3 のボクセルにリフォーマットされ、最後に 8mm のガウシアンカーネルを用いて空間的平滑化を行う。

5.2. 実験の結果と考察

5.2.1 行動データの解析結果

まず、被験者の行動系列データの解析を行った。刺激が提示されてからボタンを押すまでの反応時間 (response time: RT) の平均値を調べたところ、MEM セッションでは 3773 ミリ秒、MDP セッションでは、それぞれ 4082 ミリ秒、3774 ミリ秒、3497 ミリ秒であった。有意検定を行った結果、RT のタスク条件間での有意な差は見られなかったが、 MDP セッションでセッションが進むにつれて RT が有意に減少することが分かった ($p < 0.05$)。これは、学習が進行するにしたがって思考時間が短くなり、行動がほぼ自動的になるためであると考えられる。また、MDP セッションにおける行動のばらつきと正答率の変化を調べるために、エントロピーとオーバーラップを求めた。図 5.3の上図と下図はそれぞれ、行動系列

のエントロピーの移動平均値、および行動系列と正答系列とのオーバーラップの移動平均値であり、被験者で平均した値の時間変化を示した。移動平均の窓の大きさはどちらも 11 試行とした。図 5.3から、時間とともにオーバーラップが増加する一方で、エントロピーは減少することが分かる。したがって、正しい行動系列の学習により行動のばらつきが減少すると言える。

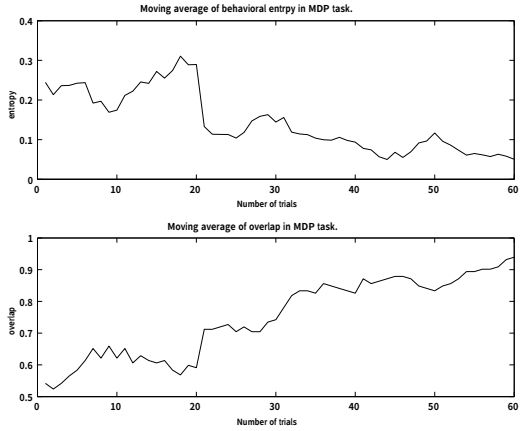

Figure 5.3 エントロピーとオーバーラップの移動平均値

5.2.2 画像データの解析結果

タスク条件間の BOLD 活動の比較には、$p < 0.005$ の有意水準でのグループランダム効果に基づき解析を行った。MEM タスクと比較して、MDP タスクで有意に活動した部位を図 5.4に示す。図 5.4から分かるように、MDP タスクでの主な活動部位は、前部 DLPF(BA46/9)、後部 DLPF(BA8)、後部頭頂葉 (BA40)、ACC(BA32) の大きく 4 つのクラスタに分けられる。表 5.1に、これらのクラスタのピーク値を取ったボクセルの統計値をまとめた。

背外側前頭前野と後部頭頂葉

MEM タスクも MDP タスクも正しいオートマトンの長さは 8 であるが、MDP 条件では、状態遷移が確率的であるため、特に学習の初期段階ではいくかの可能

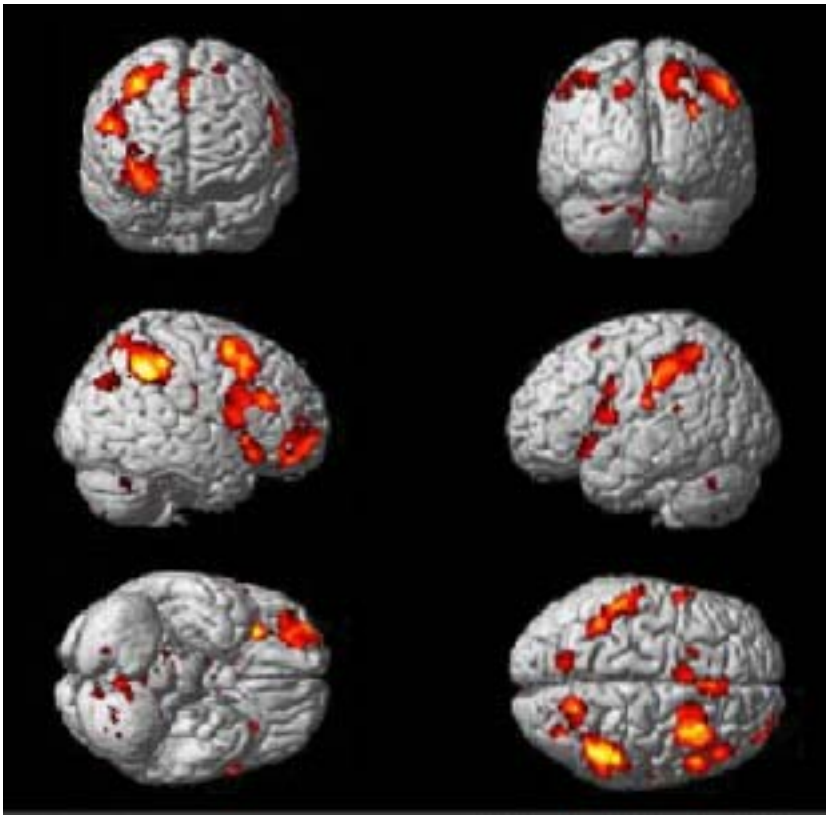

Figure 5.4 グループランダム効果解析に基づく MDP-MEM 比較時の脳活動

な系列を保持しなければならない。MDP 条件における DLPF の有意な活動は、環境モデルを表現するオートマトンの保持とその操作に関係すると考えられる。

MDP 条件では後部 DLPF と後部頭頂葉の頭頂間溝 (intraparietal sulcus: IPS) の活動が有意であった。最近の事象関連 fMRI 実験は、空間的作業記憶タスク遂行中のヒトの脳活動計測し、これらの部位が空間的情報を保持している間に持続的な活動を示すことを明らかにした [60]。さらに、これらの部位は作業記憶に保持された情報から行動に必要な情報を選択する際には活動しない。このことから、後部 DLPF と IPS は単純な情報保持機能を持つことが示唆され、本実験における環境モデルの保持に関わる部位だと考えられる。また、IPS は視覚空間的注意に重要な役割を示しており [17]、MDP タスクでの有意な活動は、行動エラーに

Table 5.1 MDP タスクで有意に活動した部位の統計値

Brain region		statistics			MNI		
Region	BA	$p_{\mathrm{corrected}}$	t-value	(Z)	x	y	z
Prefrontal cortex							
middle frontal gyrus	46/9	0.443	6.15	(4.28)	48	30	22
middle frontal gyrus	8	0.157	7.11	(4.64)	32	18	48
Parietal cortex							
inferior parietal gyrus	40	0.022	8.85	(5.16)	38	-50	40
Anterior cingulate							
cingulate gyrus	32	0.115	7.38	(4.73)	6	22	38
cingulate gyrus	32	0.529	5.96	(4.20)	-4	18	46

よって引き起こされた視覚的注意にも影響を受けていると考えられる。

一方、前部 DLPF(BA46/9) は系列情報の操作に関係すると考えられている。Pochon ら [53] は、遅延反応タスクを用いた fMRI 実験を行い、照合タスクと再現タスクという 2 種類の条件間での脳活動を比較した。照合タスクは、遅延時間中にサンプル刺激の空間的情報を保持しておき、手掛かり刺激と同じかどうかと判断するタスクである。一方、再現タスクでは、サンプル刺激の時系列情報を記憶し、遅延時間中に行動系列として保持しておかなければならない。これらの 2 条件での脳活動を比較したところ、DLPF(BA46) が再現タスクで有意に活動することが分かった。この結果から、彼らは DLPF が、作業記憶での情報の保持よりもむしろ、その情報を用いた行動系列のプランニングと準備に関わると指摘している。また、系列文字記憶タスクを用いた実験 [15] は、1 つ前の刺激を記憶するときと比べて、2 つあるいは 3 つ前までの複数の刺激を記憶するときに、DLPF(BA9) の活動が有意に増加することを示した。さらに、無記憶タスクと 1 つ前の刺激を記憶するタスクを比べた場合には活動差が見られないことから、この部位が系列情報に関わると考えられる。本実験の MDP 条件では、正しい行動系列を見つけるまでは複数の行動系列候補を準備しなければならないため、保持された環境モ

デル、すなわち状態遷移系列情報の操作が必要となる。実際に、MEM 条件の後半で、被験者が学習した系列を自動的に繰り返しているときには、この部位の活動は見られなかった。

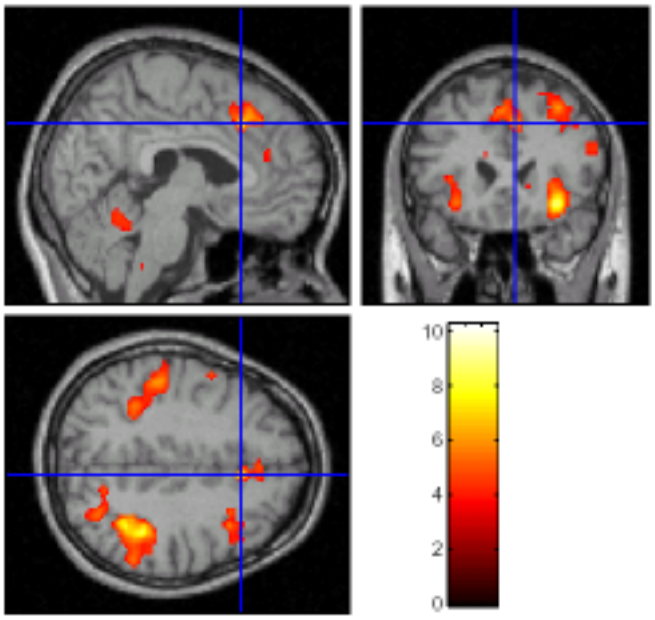

Figure 5.5 MDP-MEM 比較でみられる前部帯状回の活動増加

前部帯状回

MDP 条件と MEM 条件の比較により、MDP 条件で ACC(BA32) が有意に活動することも明らかになった。図 5.5は、ある被験者の解剖学的脳画像を標準化した画像の上に、MDP 条件で有意に活動した部分を描写したものである。図中の 2 直線の交点は MNI 座標系の (6,22,40) 点で、この点を含むクラスタが前部帯状回である。図 5.6はこの点の活動変化を各セッション毎に表示したものである。図の横軸は時間 (単位は秒) を示し、3 秒間の試行 20 回で 1 セッション、すなわち 1 セッションの長さは 90 秒である。図の縦軸は活動度を示す。図 5.6(a) の実線は、MEM セッションと 3 つの MDP セッションの計 4 セッションについて、PSTH(peri-stimulus time histgram) 法を用いて活動度の平均と分散の時間変化を

調べた結果である。PSTH 法とは、時間軸上を一定幅の小区間に分割し、各区間での活動度の試行全体にわたる平均値をその区間の代表値としてヒストグラムで表示する解析法である。4 本の線はそれぞれ、MEM セッション (赤色)、MDP1 セッション (青色)、MDP2 セッション (緑色)、MDP3 セッション (水色) での変化を示す。図 5.6(a) から、全てのセッションにおいて、時間が経つにつれて ACC の活動が減少する傾向にあり、また、MEM タスクに比べて MDP タスクで活動度が増加することが分かる。また、図 5.6(a) の破線は各セッションでの平均的な活動度を表す。このうち、MDP タスクに関するデータを取り出し、図 5.6(b) にプロットした。図 5.6(b) の実線はそれぞれ、MDP1 セッション (赤色)、MDP2 セッション (青色)、MDP3 セッション (緑色) の活動度である。図 5.6(b) から、ACC の活動は MDP セッションが進むに従って減少することが分かる。行動データから、MDP セッションの進行に伴って行動のエントロピーが減少することが示されており、ACC の活動度の減少は、学習による行動のばらつきの減少と関連していると考えられる。この解釈は、ACC が行動選択の不確実性を表現しているという仮説と一致する。

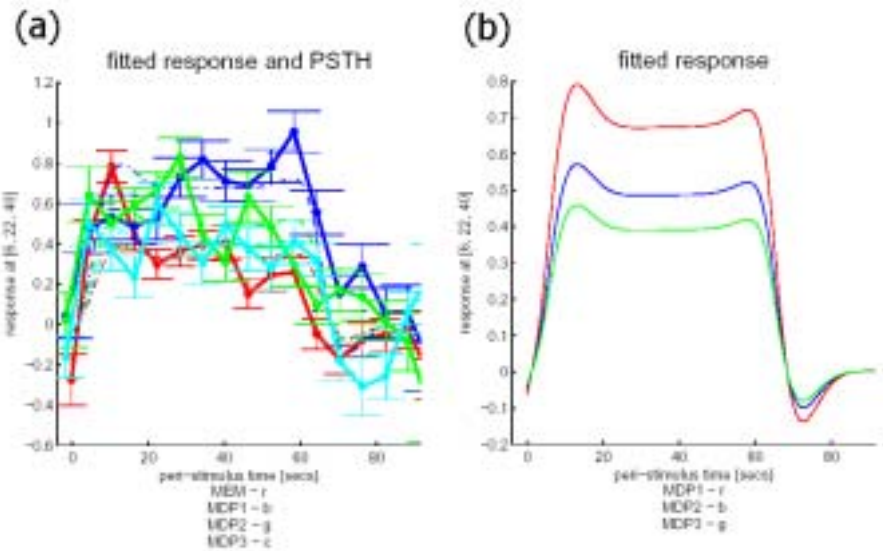

Figure 5.6 各セッションにおける前部帯状回の活動度変化

5.3. fMRI 実験のまとめと考察

本章では、第 4章で提案した脳内モデルの妥当性を検証するために、核磁気共鳴画像 (fMRI) を用いた認知心理学実験を行った。タスク条件として、強化学習タスク (MDP タスク) と系列記憶タスク (MEM タスク) の 2 種類を用意し、タスク遂行時の脳活動を比較した。脳画像データの統計解析結果から、MDP タスク時に、前部背外側前頭前野 (DLPF)、後部 DLPF、前部帯状回 (ACC)、後部頭頂葉の頭頂間溝 (IPS) での活動が有意に増加することが分かった。また、DLPF の活動は、MEM タスクの後半で見られなくなり、ACC の活動は学習が進むに従って減少する。この結果は、DLPF が環境モデルの操作と維持に、ACC が行動選択の不確実性に関わることを示唆しており、第 4章で提案した脳内モデルと矛盾しないものである。さらに、MDP タスクにおける IPS の活動は、空間的視覚注意レベルに関係するという従来の仮説からも説明できる。

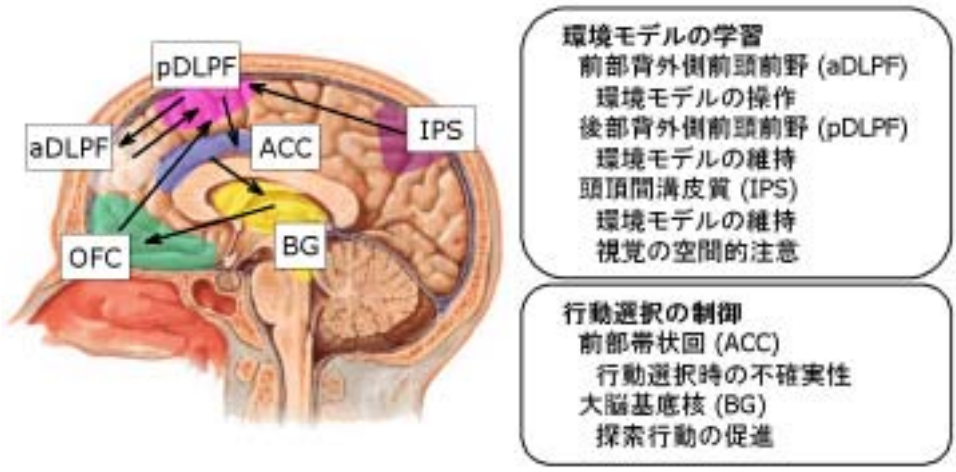

Figure 5.7 環境モデルの学習と行動選択の制御に関わる脳部位ネットワーク

第 4章で提案した脳内モデルと本実験から得られた知見を元に、環境モデルの学習と行動選択の制御に関わる脳内ネットワークを考案し、図 5.7に示した。

環境モデルの学習と保持には、主に DLPF と APF が役割を果たすと考えられる。IPS は DLPF と同時に活動することが多く、また、これらの領域間に強い直

接結合があることから、情報の保持に関わるとされている。しかし、この部位は空間的注意を司るため、提示された視覚刺激のうち必要なものに選択的注意を行うことによって情報保持の処理能力を上げる役割を持つ可能性もある。

行動選択の制御には、ACC と大脳基底核 (Basal ganglia: BG) が関わる。ACC は、大脳皮質からの指令に基づいて BG を制御することにより、行動選択を変化させる。BG は運動制御に強く関わり、損傷することによって様々な運動機能障害が起こることが知られているが、その内容は必ずしも単純ではない。例えば、淡蒼球を破壊すると、視覚的に運動を開始あるいは制御することはできるが、作業記憶に基づいた運動を行うことができなくなる。また、大脳基底核の疾患として知られるパーキンソン病患者は、外界の現象を予測して行動を起こすことができなくなる。大脳基底核によって発現される運動の多くは、外界の刺激に対する単なる反射ではなく、内的な情報に基づく予測的行動である。このような行動を行うために、大脳皮質からの情報が ACC を介して伝達されると考えられる。

Chapter 6

結論と今後の発展

6.1. まとめ

本論文では、学習による対象システムのモデル化とそのモデルを用いた予測に基づく意思決定過程に関する研究を、システム脳科学の枠組で議論した。

第 2章では、対象として非線形力学システムである低次元カオス力学系を取り上げ、その観測時系列からシステムを同定する手法を提案した。本研究では、学習対象である力学系の力学変数の一部のみが観測可能である状況を想定し、埋め込み手法を用いて実験を行った。従来しばしば用いられてきたのは遅れ座標埋め込みである。2.3.1節の実験により、NGnet に遅れ座標系でのベクトル場を学習させられることを示した。また、本手法はステムノイズおよび観測ノイズが加えられた場合においてもカオス力学系を良く学習でき、ノイズ条件下で学習した NGnet が非常に良い予測性能を持つことも示された。2.3.2節ではさらに、平滑化フィルタを用いた積分埋め込み手法を提案した。従来の遅れ座標埋め込み手法と比較したところ、新しい積分埋め込み手法の方がノイズに対してより頑強であることが実験により分かった。これは、平滑化フィルタが学習データに付加されたノイズを除去する効果を持つためだと考えられる。また、相関次元と最大リアプノフ指数が真の値とより良く一致するため、積分埋め込み手法の近似精度が高いことが分かる。さらに、この手法は必要な学習データ数が少なく、ノイズが強くなってもユニットの数が増加しないため、従来の手法より学習が早いという長所を持つ。

第 3章では、直接観測できない状態変数を持つ部分観測環境におけるモデル同定強化学習法を提案した。本研究では簡単のため、隠れ状態を多項事象であると仮定し、その確率分布を多項分布のベイズ推定によって求めた。ベイズ推定に無情報事前知識と過去の事象に対する忘却効果を導入することで、環境に対する適応性の促進をはかった。また、エージェントの行動選択のランダム性を適切に制御するために、逆温度メタパラメータの制御機構を導入した。従来手で決められていたメタパラメータを、エージェントの推定値から自動的に決める手法は新しい。本手法では、2 種類の係数によって逆温度を制御した。1 つは局所係数と呼ばれるもので、エージェントの行動価値関数のばらつき、つまりエージェントの現在の状態に基づいて決められる。もう 1 つは、環境の変化が認識されるときに全ての状態に対して探索を促す大域係数である。本手法ではさらに、エントロピーに比例した探索ボーナスも用いた。迷路探索問題へ適用した結果、本手法のエージェントは、従来の逆温度固定エージェントよりも環境への適応能力が優れていることが分かった。

次に、第 3章で議論したモデル同定強化学習法の脳内モデルを第 4章で提案した。これまでの神経生理研究や脳活動計測研究の結果から、報酬に関連した環境モデルの推定と保持には背外側前頭前野が、隠れ状態の推定には前部前頭前野が関わると想定した。また、環境変化によって引き起こされる行動選択法の変化に、前部帯状回と青斑核が重要な役割を持つと考えた。第 5章では核磁気共鳴画像を用いた認知実験を行い、提案したモデルの妥当性を検証した。単純な記憶課題とマルコフ決定過程課題を行っている間の脳活動を比較することにより、提案したモデルの一部に関して妥当性が検証された。

6.2. 今後の発展

6.2.1 コミュニケーションの発達と脳機能

本論文の第 2章では、コミュニケーション信号のカオス性を想定し、低次元カオスシステムの同定法について議論した。本手法では、複雑な時系列からもとの力学系の普遍的なダイナミクスを学習する。この学習過程は、コミュニケーション言語の発達過程、すなわち大量の発話データをもとに言語の文法を獲得する過

程と類似する。また、提案手法で用いた学習器である NGnet は、入力空間内に適当に配置されたユニットが、学習が進行するにつれてデータの出易い部分空間に移動し、不要なユニットは削除される。この、学習に伴うユニットの淘汰は、初めは無秩序に張り巡らされた脳神経系ネットワークが発達に伴ってその結合強度を変化させる現象と似ている。私は、コミュニケーションの発達過程に関する理論的枠組を提案し、その脳内モデルの構築を目指す。

コミュニケーションに関わる脳機能研究として、言語処理の脳内メカニズムを解明しようとする研究が多数行われている。しかしながら、その多くは既に獲得された言語の処理過程を明らかにしようとしたもので、新たな言語の学習や発達過程を通したコミュニケーション能力の変化について調べたものは少ない。私は、子供のコミュニケーション能力の発達的研究や、自閉症などの疾患による能力低下といった神経心理学的研究など、広い視野からコミュニケーションを捉え、その計算モデルを検討する。また、時系列解析の手法を用いて生成された言語の性質を定量的に評価することにより、言語の複雑度と脳機能の関係を明らかにすることが期待できる。

6.2.2 連続システムの同定と強化学習

実問題における環境は連続で非線形である。しかし、特に部分観測のような複雑な環境下で連続世界を仮定することは困難であり、強化学習問題として取り扱うにはモデルの簡易性を仮定せざるを得ない。そのため、本論文の第 3章で述べた手法は、対象となる環境が離散的であることを仮定して導出された。連続環境におけるモデル同定強化学習法については、現在研究を行っている。この手法では、環境を線形ガウスシステムと仮定しオンライン変分ベイズ法によって学習する。今後、この研究の中で逆温度メタパラメータの脳内機能についてより詳しく論じる予定である。この新しい手法は、連続空間という現実的な問題を扱っており、本論文の第 2章で述べた連続システムの同定法との関連もより強いものとなる。また、非線形システムへの発展に関しても、現在研究を実施中である。

6.2.3 隠れ変数の推定と脳機能局在

第 4章で提案した脳の情報処理モデルでは、隠れ状態の推定に APF が関係していると想定している。APF は PFC の中でも最も前方に位置し、他の霊長類と比べてヒトでの発達が最も顕著な部位である。そのため、ヒト特有の高次情報処理に関わると予想できるが、その具体的な機能は未だ明らかになっていない。しかし、いくつかの研究結果から、APF がコミュニケーションの脳内過程に大きく関わっていると考えられる。相手の心を読むという過程で APF が活動し、相手の心を理解できない自閉症の子供は、APF の活動が見られないことが知られている。また、APF は最も発達が遅い脳部位である一方で、加齢にともなう脳活動レベルの低下は他の脳の部位よりも顕著である。また、APF は局所的な破壊や電位計測が困難なため、サルを用いた障害・計測実験が事実上不可能である。ヒトの非侵襲計測実験を用いた APF の機能理解への試みは非常に有意義である。現在は、この APF の機能解明を目指した認知科学実験を行っている。

また、fMRI 実験では、各機能に関わる活動部位は特定できるものの、その時間変化や事象関連活動を調べることは難しい。そこで、時間分解能の高い脳磁図を用いて強化学習タスク遂行中の事象関連活動の計測を行うことを予定している。脳活動の時間変化を詳細に計測することによって、複数の部位間の情報の流れや、例えば行動ルールを切替えた時などに一過性に発生する脳活動を明らかにすることが期待される。

参考文献

[1] Abarbanel, H.I.D. (1996). *Analysis of observed chaotic data*, Springer-Verlag, New York.

[2] Aston-Jones, G., Chiang, C., and Alexinsky, T. (1991). Discharge of noradrenergic locus coeruleus neurons in behaving rats and monkeys suggests a role in vigilance. *Progress in Brain Research*, **88**, 501-520.

[3] Aston-Jones, G., Rajkowski, J., Kubiak, P., and Alexinsky, T. (1994). Locus coeruleus neurons in monkey are selectively activated by attended cues in a vigilance task. *The Journal of Neuroscience*, **14**, 4467-4480.

[4] Barbas, H., and Pandya, D.N. (1989). Architecture and intrinsic connections of the prefrontal cortex in the rhesus monkey. *The Journal of comparative neurology*, **286**, 353-375.

[5] Barto, A.G., Sutton, R.S., and Anderson, C.W. (1983). Neuronlike elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, **13**, 835-846.

[6] Barto, A.G. (1995). Adaptive critics and the basal ganglia. In J.C. Houk, J.L. Davis, and D.G. Beiser, Eds., *Models of Information Processing in the Basal Ganglia.* Cambridge, MA: MIT Press, pp. 215-232.

[7] Berns, G.S., Cohen, J.D., and Mintun, M.A. (1997). Brain regions responsive to novelty in the absence of awareness. *Science*, **276**, 1272-1275.

[8] Botvinick, M., Nystrom, L.E., Fissell, K., Carter, C.S., and Cohen, J.D. (1999). Conflict monitoring versus selection-for-action in anterior cingulate cortex. *Nature*, **402**, 179-181.

[9] Brafman, R.I., and Tennenholtz, M. (2001). R-max: a general polynomial time algorithm for near-optimal reinforcement learning, In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, pp. 953-958.

[10] Braver, T.S., Barch, D.M., Gray, J.R., Molfese, D.J., and Snyder, A. (2001). Anterior cingulate cortex and response conflict: effects of frequency, inhibition and errors. *Cerebral Cortex*, **11**, 825-836.

[11] Bush, G., Vogt, B.A., Holmes, J., Dale, A.M., Greve, D., Jenike, M.A., and Rosen, B.R. (2002). Dorsal anterior cingulate cortex: A role in reward-based decision making. *Proceedings of the National Academy of Sciences, USA*, **99**, 507-512.

[12] Carter, C.S., Braver, T.S., Barch, D.M., Botvinick, M.M. Noll, D., and Cohen, J.D. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science*, **280**, 747-749.

[13] Casdagli,M., Eubank, S., Farmer, J.D., and Gibson, J. (1991). State space reconstruction in the presence of noise. *Physica D*, **51**, 52-98.

[14] Christie, M.J., Williams, J.T., and North, R.A. (1989). Electrical coupling synchronizes subthreshold activity in locus coeruleus neurons in vitro from neonatal rat. *The Journal of Neuroscience*, **9**, 3584-3589.

[15] Cohen, J.D., Perlstein, W.M., Braver, T.S., Nystrom, L.E., Noll, D.C., Jonides, J., and Smith, E.E. (1997). Temporal dynamics of brain activation during a working memory task. *Nature*, **386**, 604-608.

[16] Cohen, J.D., Botvinick, M., and Carter, C.S. (2000). Anterior cingulate and prefrontal cortex: who's in control?. *Nature neuroscience*, **3**, 421-423.

[17] Corbetta, M., Kincade, J.M., Ollinger, J.M., McAvoy, M.P., and Shulman, G.L. (2000). Voluntary orienting is dissociated from target detection in human posterior parietal cortex. *Nature Neuroscience*, **3**, 284-291.

[18] Dayan, P., and Sejnowski, T.J. (1996). Exploration bonuses and dual control. *Machine Learning*, **25**, 5-22.

[19] Dearden, R., Friedman, N., and Andre, D. (1999). Model based Bayesian exploration. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, San Francisco, CA: Morgan Kaufman, 150-159.

[20] Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society B*, **39**, 1-22.

[21] Doya, K. (2000). Reinforcement learning in continuous time and space, *Neural Computation*, **12**, 219-245.

[22] Doya, K. (2000). Metalearning, neuromodulation, and emotion. In *Affective Minds* (eds. G.Hatano, N.Okada & H.Takabe), Elsevier Science, 101-104.

[23] Eckmann, J.-P., Kamphorst, S.O., Ruelle, D., and Ciliberto, S. (1986). Lyapunov exponents from time series. *Physical Review A*, **34**, 4971-4979.

[24] Elliott, R., Dolan, R.J., and Frith, C.D. (2000). Dissociable functions in the medial and lateral orbitofrontal cortex: evidence from human neuroimaging studies. *Cerebral Cortex*, **10**, 308-317.

[25] Fe'ldbaum, A.A. (1965). *Optimal Control Systems*, New York, NY: Academic Press.

[26] Foote, S.L., Bloom, F.E., and Aston-Jones, G. (1983). Nucleus locus coeruleus: new evidence of anatomical and physiological specificity. *Physiological Reviews*, **63**, 844-914.

[27] Gehring, W.J., Goss, B., Coles, M.G.H., Meyer, D.E., and Donchin, E. (1993). A neural system for error detection and compensation. *Psychological Science*, **4**, 385-390.

[28] Gonzales, C., and Chesselet, M.-F. (1990). Amygdalonigral pathway: an anterograde study in the rat with *Phaseolus vulgaris* leucoagglutinin (PHA-L). *The Journal of comparative neurology*, **297**, 182-200.

[29] Grassberger, P., and Procaccia, I. (1983). Characterization of strange attractors. *Physical Review Letters*, **50**, 346-349.

[30] Heckerman, D. (1999). A tutorial on learning with Bayesian networks. In *Learning in Graphical Models* (ed. M.I.Jordan), Cambridge, MA: MIT Press, 301-354.

[31] Hoshi, E., Shima, K, and Tanji, J. (2000). Neuronal activity in the primate prefrontal cortex in the process of motor selection based on two behavioral rules. *Journal of Neurophysiology*, **83**, 2355-2373.

[32] 石井 信, 佐藤 雅昭. (1999). 正規化ガウス関数ネットワーク, Mixture of experts と EM アルゴリズム. 日本神経回路学会誌, **6(1)**, 30-40.

[33] Ishii, S. (2001). Control of exploration-exploitation balance in reinforcement learning, *CREST Workshop on Metalearning and Neuromodulation*, (Kyoto, Apr., 2001).

[34] Ishii, S., and Sato, M. (2001). Reconstruction of chaotic dynamics based on on-line EM algorithm. *Neural Networks*, **14**(9), 1239-1256.

[35] Ishimaru, M., and Williams, J.T. (1996). Synchronous activity in locus coeruleus results from dendritic interactions in pericoerulear regions. *The Journal of Neuroscience*, **16**, 5196-5204.

[36] Jenkins, I.H., Brooks, D.J., Nixon, P.D., Frackowiak, R.S.J., and Passingham, R.E. (1994). Motor sequence learning: A study with Positron Emission Tomography. *The Journal of Neuroscience*, **14**, 3775-3790.

[37] Jueptner, M., Stephan, K.M., Frith, C.D., Brooks, D.J., Frackowiak, R.S., and Passingham, R.E. (1997). Anatomy of motor learning. I. Frontal cortex and attention to action. *The Journal of Neurophysiology*, **77**, 1313-1324.

[38] Kaelbling, L. (1993). *Learning in Embedded Systems*, Cambridge, MA: MIT Press.

[39] Kaelbling, L.P., Littman, M.L., and Cassandra, A.R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, **101**, 99-134.

[40] Kearns, M., and Singh, S. (1998). Near-optimal performance for reinforcement learning in polynomial time. In *Proceedings of the 15th International Conference on Machine Learning*, San Mateo, CA: Morgan Kaufmann, 260-268.

[41] Kennel, M.B., Brown, R., and Abarbanel, H.I.D. (1992). Determining minimum embedding dimension using a geometrical construction. *Physical Review A*, **45**, 3403-3411.

[42] Koechlin, E., Corrado, G., Pietrini, P., and Grafman, J. (2000). Dissociating the role of the medial and lateral anterior prefrontal cortex in human planning. *Proceedings of the National Academy of Sciences, USA*, **97**, 7651-7656.

[43] Leon, M.I., and Shadlen, M.N. (1999). Effect of expected reward magnitude on the response of neurons in the dorsolateral prefrontal cortex of the macaque. *Neuron*, **24**, 415-425.

[44] Matsumoto, T., Nakajima, Y., Hamagishi, H., Sugi, J., and Saito, M. (1998). From data to nonlinear dynamics: A hierarchical Bayes approach with neural nets. *Neural Networks for Siginal Processing VIII*, 333-342, IEEE, New York.

[45] Matsuno, Y., Yamazaki, T., Matsuda, J., and Ishii, S. (2001). A multi-agent reinforcement learning method for a partially-observable competitive game.

In *Proceedings of the Fifth International Conference on Autonomous Agents*, ACM, 39-40.

[46] Meunier, M., Bachevalier, J., and Mishkin, M. (1997). Effects of orbital frontal and anterior cingulate lesions on object and spatial memory in rhesus monkeys. *Neuropsychologia*, **35**, 999-1015.

[47] Moody, J., and Darken, C.J. (1989). Fast learning in networks of locally-tuned processing units. *Neural Computation*, **1**, 281-294.

[48] Moore, A.W., and Atkeson, C.G. (1993). Prioritized sweeping: reinforcement learning with less data and less real time. *Machine Learning*, **13**, 103-130.

[49] Morrison, J., and Foote, S. (1986). Noradrenergic and serotonergic innervation of cortical, thalamic and tectal visual structures in old and new world monkeys. *The Journal of Comparative Neurology*, **243**, 117-128.

[50] O'Doherty, J., Kringelbach, M.L., Rolls, E.T., Hornak, J., and Andrews, C. (2001). Abstract reward and punishment representations in the human orbitofrontal cortex. *Nature Neuroscience*, **4**, 95-102.

[51] Pennartz, C.M., Groenewegen, H.J., and Lopez de Silva, F.H. (1994). The nucleus accumbens as a complex of functionally distinct neuronal ensembles: an integration of behavioural, electrophysiological and anatomical data. *Progress in Neurobiology*, **42**, 719-761.

[52] Picard, N., and Strick, P.L. (1996). Motor areas of the medial wall: a review of their location and unctional activation. *Cerebral Cortex*, **6**, 342-353.

[53] Pochon, JB., Levy, R., Poline, JB., Crozier, S., Lehericy, S., Pillon, B., Deweer, B., Le Bihan, D., and Dubois, B. (2001). The role of dorsolateral prefrontal cortex in the preparation of forthcoming actions: an fMRI study. *Cerebral Cortex*, **11**, 260-266.

[54] Posner, M.I., and Raichle, M. (1996). *Images of Mind (Revised)*. Washington, DC: Scientific American Books.

[55] Rajkowski, J., Lu, W., Zhu, Y. Cohen, J., and Aston-Jones, G. (2000). Prominent projections from the anterior cingulate cortex to the locus coeruleus in rhesus monkey. *Society of Neuroscience Abstract*, **26**, 2230.

[56] Rao, S.C., Rainer, G., and Miller, E.K. (1997). Integration of what and where in the primate prefrontal cortex. *Science*, **276**, 821-824.

[57] Rizolatti, G., Fadiga, L., Gallese, V., and Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Reseach*, **3**, 131-141.

[58] Rolls. E.T., Hornak, J., Wade, D., and McGrath, J. (1994). Emotion-related learning in patients with social and emotional changes associated with frontal lobe damage. *Journal of Neurology, Neurosurgery, and Psychiatry*, **57**, 1518-1524.

[59] Rolls, E.T. (1996). The orbitofrontal cortex. *Philosophical Transactions of the Royal Society of London. Series B: Biological sciences*, **351**, 1433-1443.

[60] Rowe, J.B., Toni, I., Josephs, O., Frackowiak, R.S.J. and Passingham, R.E. (2000). The prefrontal cortex: response selection or maintenance within working memory?. *Science*, **288**, 1656-1660.

[61] Sato, M., Murakami, Y., and Joe, K. (1990). Learning chaotic dynamics by recurrent neural networks. *Proc. of International Conference on Fuzzy Logic & Neural Networks*, 601-605.

[62] Sato, M. (1990). A learning algorithm to teach spatiotemporal patterns to recurrent neural networks. *Biological Cybernetics*, **62**, 259-263.

[63] Sato, M., and Murakami, Y. (1991). Learning nonlinear dynamics by recurrent neural networks. *Proc. of Symposium on Some Problems on the Theory of Dynamical Systems in Applied Science*, 49-63, Singapore, World Scientific.

[64] Sato, M., and Ishii, S. (1999). Reinforcement learning based on on-line EM algorithm. *Advances in Neural Information Processing Systems 11*, 1052-1058, MIT Press, Reading.

[65] Sato, M., and Ishii, S. (2000). On-line EM algorithm for the normalized Gaussian network. *Neural Computation*, **12**(2), 407-432.

[66] Sato, M. (2001). On-line model selection based on the variational Bayes. *Neural Computation*, **13**, 1649-1681.

[67] Sauer, T., Yorke, J.A., and Casdagli, M. (1991). Embedology. *Journal of Statistical Physics*, **65**. 579-616.

[68] Schoenbaum, G., Chiba, A.A., and Gallagher, M. (1998). Orbitofrontal cortex and basolateral amygdala encode expected outcomes during learning. *Nature Neuroscience*, **1**, 155-159.

[69] Schultz, W., Dayan, P., and Montague, R.P. (1997). A neural substrate of prediction and reward. *Science*, **275**, 1593-1599.

[70] Schultz, W. (1998). Predictive reward signal of dopamine neurons. *The Journal of Neurophysiology*, **80**, 1-27.

[71] Servan-Schreiber, D., Printz, H., and Cohen, J.D. (1990). A network model if catecholamine effects: gain, signal-to-noise ratio, and behavior. *Science*, **249**, 892-895.

[72] Shima, K., and Tanji, J. (1998). Role for cingulate motor area cells in voluntary movement selection based on reward. *Science*, **282**, 1335-1338.

[73] Singh, S.P., Jaakkola, T., and Jordan, M.I. (1994). Learning without state-estimation in partially observable Markovian decision processes. In *Proceedings of the Eleventh International Conference on Machine Learning*, San Francisco, CA: Morgan Kaufmann, 284-292.

[74] Stern, C.E., Corkin, S., Gonzalez, R.G., Guimaraes, A.R., Baker, J.R., Jennings, P.J., Carr, C.A., Sugiura, R.M., Vedantham, V., and Rosen, B.R. (1996). The hippocampal formation participates in novel picture encoding: evidence from functional magnetic resonance imaging. *Proceedings of the National Academy of Sciences, USA*, **93**, 8660-8665.

[75] Strange, B.A., Henson, R.N.A., Friston, K.J., and Dolan, R.J. (2001). Anterior prefrontal cortex mediates rule learning in humans. *Cerebral Cortex*, **11**, 1040-1046.

[76] Sutton, R.S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, **3**, 9-44.

[77] Sutton, R.S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Machine Learning: Proceeding of the Seventh International Conference*, pp. 216-224.

[78] Sutton, R.S., and Barto, A.G. (1998). *Reinforcement Learning: An Introduction*, Cambridge, MA: MIT Press.

[79] Takens, F. (1981). Detecting strange attractors in fluid turbulence. *Dynamical Systems and Turbulence*, Springer-Verlag, Berlin.

[80] Tanji, J., and Hoshi, E. (2001). Behavioral planning in the prefrontal cortex. *Current Opinion in Neurobiology*, **11**, 164-170.

[81] Taylor, J.R. and Robbins, T.W. (1986). 6-Hydroxydopamine lesions of the nucleus accumbens, but not of the caudate nucleus, attenuate enhanced responding with reward-related stimuli produced by intra-accumbens d-amphetamine. *Psychopharmacology*, **90**, 390-397.

[82] Thrun, S.B. (1992). The role of exploration in learning control. In *Handbook of Intelligent Control: Neural, Fuzzy and Adaptive Approaches*, Frorence, KY: Van Nostrand Reinhold, pp. 527-559.

[83] Tokuda, I. (1993). Deterministic prediction and speech signals of the Japanese vowel /a/. Master Thesis for University of Tsukuba.

[84] Tulving, E., Markowitsch, H.J., Craik, F.E., Habib, R., and Houle, S. (1996). Novelty and familiarity activations in PET studies of memory encoding and retrieval. *Cerebral Cortex*, **6**, 71-79

[85] Usher, M., Cohen, J.D., Servan-Schreiber, D., Rajkowski, J., and Aston-Jones, G. (1999). The role of locus coeruleus in the regulation of cognitive performance. *Science*, **283**, 549-554.

[86] Vogt, B.A., Finch, D.M., and Olson, C.R. (1992). Functional heterogeneity in cingulate cortex: the anterior executive and posterior evaluative regions. *Cerebral Cortex*, **2**, 435-443.

[87] Waelti, P., Dickinson, A., and Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature*, **412**, 43-48.

[88] Watanabe, M. (1996). Reward expectancy in primate prefrontal neurons. *Nature*, **382**, 629-632.

[89] Watkins, C.J.C.H., and Dayan, P. (1992). Technical note: Q-learning. *Machine Learning*, **8**, 279-292.

[90] Wilson, F.A.W., and Rolls, E.T. (1993). The effects of stimulus novelty and familiarity on neuronal activity in the amygdala of monkeys performing recognition memory tasks. *Experimental Brain Research*, **93**, 367-382.

[91] Xu, L., Jordan, M.I., and Hinton, G.E. (1995). An alternative model for mixtures of experts. *Advances in Neural Information Processing Systems 7*, 633-640, MIT Press, Reading.

[92] Yim, C.Y., and Mogenson, G.J. (1989). Low doses of accumbens dopamine modulate amygdala suppression of spontaneous exploratory activity in rats. *Brain Research*, **477**, 202-210

[93] 吉本 潤一郎, 石井 信, 佐藤 雅昭. (2000). オンライン EM アルゴリズムによる強化学習法の acrobot 制御への応用, 電子情報通信学会論文誌, J83-D-II (3), 1024-1033.

付録

A. 業績リスト

- 論文

  1. 吉田 和子, 石井 信, 佐藤 雅昭: オンライン EM アルゴリズムによるカオス力学系の学習と耐ノイズ性, 電子情報通信学会論文誌, J83-A(1), 28-37, (2000).
  2. S.Ishii, W.Yoshida and J.Yoshimoto: Control of exploitation-exploration meta-parameter in reinforcement learning, *Neural Networks*, **15**(4-6), 665-687, (2002).

- 国際会議

  1. W.Yoshida, S.Ishii and M.Sato: Reconstruction of chaotic dynamics and robustness to noise with on-line EM algorithm, 1999 IEEE International Conference on Systems, Man and Cybernetics (IEEE-SMC), I, 414-419, New York: IEEE (Tokyo, Sep., 1999).
  2. W.Yoshida, S.Ishii and M.Sato: Approximating discrete mapping of chaotic dynamical system based on on-line EM algorithm. 1999 International Conference on Neural Information Processing (ICONIP), III, 1010-1016, New York: IEEE (Perth, Nov., 1999).
  3. S.Ishii, J.Yoshimoto, W.Yoshida and M.Sato: On-line EM algorithm and its applications. In *Emerging Knowledge Engineering and Connectionist-Based Information Systems*, 17-20, Dunedin: University of Otago.

4. W.Yoshida, S.Ishii and M.Sato: Reconstruction of chaotic dynamics using a noise-robust embedding method, 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), I, 181-184 (Istanbul, Jun., 2000).
5. W. Yoshida and S. Ishii: A model-based reinforcement learning: a computational model and an fMRI study, 11th European Symposium on Artificial Neural Networks (ESANN), to appear (Bruges, Apr., 2003).
6. W. Yoshida and S. Ishii: A possible function of anterior prefrontal cortex in a model-based reinforcement learning: a computational model and an fMRI study, The Annual Computational Neuroscience Meeting (CNS), submitted.

- その他
  - 再編集論文
    1. W.Yoshida, S.Ishii and M.Sato: Learning chaotic dynamics under noise with on-line EM algorithm. *Electronics and Communications in Japan (PART III: Fundamental Electronic Science)*, **84**(6), 23-31, (2001).
  - 研究会報告
    1. 吉田 和子, 石井 信, 佐藤 雅昭. オンライン EM アルゴリズムを用いたカオス力学系の再構成. 電子情報通信学会技術研究報告, NLP98-71, 9-16, (奈良, Nov., 1998).
    2. 吉田 和子, 石井 信, 佐藤 雅昭. 正規化ガウス関数ネットワークと EM アルゴリズムによるカオス力学系再構成. 電子情報通信学会技術研究報告, NC98-82, 33-40, (札幌, Feb., 1999).
    3. 吉田 和子, 石井 信, 佐藤 雅昭. 積分埋め込みを用いたニューラルネットによるカオス力学系の再構成. 電子情報通信学会技術研究報告, NC99-120, 21-28, (東京, Mar., 2000).

4. 吉田 和子, 石井 信, 佐藤 雅昭. 2 種の平滑化埋め込みを用いたニューラルネットによるカオス力学系の再構成. 電子情報通信学会技術研究報告, NLP99-171, 75-82, (東京, Mar., 2000).
5. 吉田 和子, 石井 信. 強化学習における exploration-exploitation 問題の制御. 電子情報通信学会技術研究報告, NC2001-28, 41-48, (沖縄, Jun., 2001).

– ワークショップ等

1. 吉田 和子, 石井 信, 佐藤 雅昭. オンライン EM アルゴリズムによるカオス力学系再構成と耐ノイズ性. 第 12 回回路とシステム軽井沢ワークショップ, 175-180, (軽井沢, Apr., 1999).
2. 吉田 和子. NGnet によるカオス力学系再構成. 第 1 回長岡シンポジウム, 27-37, (長岡, Mar., 2000).

– その他

1. 吉田 和子, 石井 信, 佐藤 雅昭. 平滑化埋め込みを用いたニューラルネットによるカオス力学系の再構成. 第 23 回日本神経科学大会・第 10 回日本神経回路学会 合同大会, (横浜, Sep., 2000).
2. 吉田 和子, 石井 信. 強化学習における環境の同定と行動に関する注意. 「脳を創る」シンポジウム, (東京, Jun., 2001).
3. 吉田 和子, 吉本 潤一郎, 石井 信. 強化学習における注意機構. 脳と心のメカニズム第 2 回冬のワークショップ, (ルスツ, Jan., 2002).
4. 吉田 和子, 石井 信. 強化学習における注意とその脳内メカニズム. 「脳を創る」シンポジウム, (東京, Jun., 2002).
5. 吉本 潤一郎, 石井 信, 吉田 和子, 佐藤 雅昭. (2003). オンライン変分ベイズ法による部分観測環境の同定と強化学習への応用. 脳と心のメカニズム第 3 回冬のワークショップ, (ルスツ, Jan., 2003).