学位請求論文

強化学習に基づく知能システム

価値体系を利用したパターン処理型知能マシンの検討

工学系研究科電子工学専攻　博士課程
指導教官　岡部洋一　教授
９７０８９　　山川宏

目次

序　文

人間の行うような知的な作業を機械で実現しようとする試みはそれ自体大きな夢であるとともに、脳機構の理解においても有益である。この試みはコンピュータ技術の発展により、まず数値処理において非常な成功を納め、次にはより一般的なシンボルを処理対象として人工知能の領域を形作った。

しかしながら、本来パターン的である我々の世界に対するシンボル化は一意に決定できないので、シンボル型の処理ではその方法は目的毎にシステムの製作者が検討する必要があるという限界が存在する。このことは経験から得られるパターン情報を直接に処理する能力が欠けていること示している。そのため、この種のシステムは外界から新たな情報を取り込んで製作者の意図を越える発展的な動作を行うことは無い。またシンボル化されたデータは表現された時点で理解の仕方が確定しているので、別の観点から観るような柔軟性はほとんど期待できない。　一方、今世紀半ばに生物の神経細胞を手本とする基本的には可塑性を備えたしきい値素子により構成されるるニューラルネットワークがパターン的な処理を行うことができる回路として出現した。しかし、ニューラルネットワークに関する研究は主に入力に近い認識の分野と、出力に近い制御の分野に対して応用が為されており、判断や思考等の高次の知識処理に関してはまだ研究の初期段階である。

高次の機能の実現にはやはり生物を参考にするのが有力な方法である。この場合入力と出力を合わせ持った動作主体が環境と相互作用する系を考える必要があると思われる(Okabe, 1986, 1988)。生物の行っている作業は形式的に「ある環境に対して入力と出力を合わせ持ったシステムがその出力を操作することによってある評価関数を最大化する作業」として記述できる。ここで重要なのは Widrow et al (1973) が指摘しているように、現実世界では評価関数は適切な出力を教える teacher としては存在せず、単に結果の善し悪しのみを批判する critic として存在する点である。これらの課題は強化学習として知られており国内では中野ら(1987)によるアソシアトロンを用いた研究等(佐賀 et al, 1989)が行われている。また、強化シグナルを適応的とした研究(Barto et al, 1981: Barto et al, 1983: Sutton, 1988: Jameson, 1990)や、teacher signal を得るため(credit assignment problem)に環境に対するモデルを自己形成する研究等(Werbos, 1987: Werbos,1989: Shmidhuber, 1989, 1990)やその他関連の研究(Sutton et al, 1981: Klopf, 1986, 1987: Morgan, 1990: Muro, 1986: Williams et al, 1989)が為されている。また Pavlov (1927) により発見された古典的条件付けをニューロ

ンモデルで説明する研究(Barto et al, 1982: Tesauro, 1986)も関連が大きい。

これらの経緯をふまえ本論文では環境からの入力に応じて外界に対して行動出力を行なう強化学習を用いたモジュールを持つ動作主体において、注意のポイントを与えるための重要性を含む価値体系を構築するためのモジュールと、世界モデルを構築するためのモジュールを設置したことを特徴とするパターン処理型知能マシンの提案を行なった。そして、このモデルが神経回路により構築できる可能性があることを確認し。さらに、その部分システムが環境と相互作用することにより価値体系を構築し得ることを計算機シミュレーションにより示した。

第１章

研究の動機と立場

1-1. 研究の動機とテーマの設定

「人は物質のみによっては説明できないのだろうか？」この疑問は少なくとも古代ギリシアにまでも遡ることができる根源的な問いかけである。この疑問に答えようとする観念的な試みは多くの偉大な先人達によって行われたが、今だ明確な答えは得られていない。私の最も基本的な興味も正にここにある。

過去の試みに比べて現代に生きる我々がこの問題に立ち向かうとき有利になった点を指摘する。まず第１にコンピュータの進歩により人間の心的動作の一部は機械によって実現することが可能である事が証明されたこと。第２には生物学の発達により心の動きが物質レベルである程度理解できたことである。

第１の変化のおかげで以前に比べて人間特有の機能や能力の範囲が狭められ心の本質が次第に明確になりつつある。しかしながら現状のコンピュータの能力は、まだ人間と比較できるレベルにはなく、どこまでが物質で説明できるのかを議論するには不十分である。そこで、今後の研究によりコンピュータの能力をできるだけ人間に近付けることが必要である。この方向で研究を進めることにより、明確な動作原理のみによって構成されたシステムが人間と全く区別がつかない動作を行うことができたなら、人間は物質のみで説明できることが証明される。逆にそのような究極の人間的システムができない場合は、結論は明かではないが、ある段階で物質による説明の限界が判るかも知れないし、またこの限界が新たな科学的パラダイムへの足掛りとなる可能性をもつ。であるから、この様な視点に立って研究を進めるのは非常に有意義である。この際に先に述べたように近年急速に蓄積されつつある生物学の知見を取り込むのは非常な助けとなるだろう。

以上のような探求的な興味以外にもコンピュータの能力を人間のそれに近付けるという研究には多くの興味がある。生物学的な視点としてはコンピュータの構成的な方法の成果から逆に脳の理解を深める効果が期待され、これが医学に適用されることにより精神病や

ぼけ等の治療への助けとなることが期待される。さらに、例えば記憶能力の改善等のこれまで以上に脳の能力を活用する研究や、脳の機能を考慮した社会心理学の研究などの可能性も開かれる。工学的には現在のコンピュータが人間に比べて不得意な様々な能力を開拓することである。これらの興味があいまって、近年ニューラルネットワーク周辺の研究分野では工学、心理学、生物学、医学などの交流が盛んになっている。

以上説明したような、現在の人類の知識レベルに鑑みて、ニューラルネットワークを利用してより人間に近い情報処理マシンを開発するという本研究の基本的コンセプトは大変有意義である。

コンピュータを人間に近付けるという基本的なテーマを決定しても、現在のコンピュータと人間は非常にかけ離れているため、より焦点を絞り込む必要がある。

現在の多くの研究が入出力に近い認識(福島,1989)や制御のレベルに注目しているのは、意味が明確な両端から徐々に中枢へと攻め込む方が研究を行ない易いからである。しかし先に述べたように私の基本的な興味は人間の能力を唯物的に説明することができるかということなので、明かに物質的に理解ができる末端部周辺にはあまり興味がなく、思考、判断などを伴う中枢部に興味がある。

そこで私は始めから中枢部を取り扱う少々大胆な研究方針をとった。初期の目標として「自発的に思考するマシン」の開発を検討した。つまり外部から何の命令がなくても自発的に思考し、最終的にはそれが自発的な行動に結び付くようなシステムを構築することである。

この目標を実現するためには、次の三つの研究が必要であると考えた。

1) 動作主体が環境に関する知識を貯える方法を検討する。
2) 思考に対する操作を定義する。
3) 思考を行うべき知識を選択する基準を設ける。

思考を行なうために1),2)の研究が必要であるのは当然である。Fig.1-1に自発的に思考を行なうシステムのデータの流れを示した。環境から取り入れた情報に対して思考を促す始動信号が必要であるが、この際に何を対象として思考を行なうかを決定する重要度を表す基準が必要となる、我々はこれを価値観の一つであると考えた。なお、1)の知識を貯えるデータの形式については、現実世界にマッチするパターン型であることを前提とし、情報の本質がそこに含まれる相関であると考え、2)の操作は相関の強調であると考えた。これらの点に関しては後に第２章で詳しく議論する。

1-2. 人工知能(Artficial intelligence)の限界

知能システムを考えるときに、その先進的な役割を果たしたAIに触れておく必要があるだろう。コンピュータの初期の歴史は計算機としてしての歴史であった。これは言ってみれば人間が考えた概念のうちで最も抽象化の進んだ数の概念を取り扱っている。これに対しAIの試みは、より一般的な概念をコンピュータで取り扱おうとする試みである。この方法は限られた世界や特殊な用途ではかなりの成功をおさめたが、日頃人間が行うような常識的な判断を苦手としている。この、困難を乗り越えるために知識ベースを利用する方法などもあるが、この種の手法では知識ベースをいかに造るかが問題である。なぜなら神でない我々は完全な知識を持っているわけではなく、無数の知識を全て取り扱ってしまうと情報量と処理コストが大きくなり過ぎるのである。AIの研究において、「良い成果は良い例題を選ぶことによって得られる」、という声を聞くのは、この研究状況をよく表していると言えよう。つまり、基本的にAIの手法が成功する為の条件は、処理すべき問題とそれを解決するための知識が、問題解決に適した形式で論理的に記述されているかどうかなのである。そういう視点から見ると、AIの研究は世の中のあらゆるモノがどこまで論理的に記述できるかを追及し、あわ良くば世の中の全てを記述し得る論理的記述形式を捜し出そうと日夜邁進している学問のように見える。しかし、この最後の企ての実現はほとんど不可能だと思われる。結局のところAIは論理によって記述されたデータを処理する推論マシンである。

次に、私のテーマである人間により近い知能システムをAIで作ろうとする場合にどの様な困難があるのかを指摘しておく。

第１にAIが取り扱う情報が本質的に論理的データで、意味を持つ最小単位が決っており、その単位毎にあらかじめ決められた属性がある。データを分解しても、その最小単位を分解しない限りは意味が失われることはないが、その最小単位を分解したとたんに全く意味を失う。一方、現実世界を記述するパターンデータはあらかじめ与えられた最小単位や属性がなく、その内部に含まれる相関のみに本質的な意味を持つ。分解操作を行うと次第に意味を失い、完全に分割された時に完全に意味を失う。論理データの例としてはデータベース上の数字を想像すれば良いだろう、物品フィールド上では"1"という数字は食料品を表すかも知れないが、価格フィールドではその値段を表すのである。一方パターンデータは例えば視野に黒いものが見えてもそれがカラスなのか黒い画用紙なのか単に黒いなと思うのか意味付けは多様である。

この様に、データの記述方法が非常に異なるにも関わらずAIや計算機が現実問題の解決に役に立つのは、他でもない人間がこのこの２種類の間の変換作業を請け負っているからである。本来分割されていない世界を人間がその目的に合った方法で分割することにより

AIに教え込んでいるので、AI自身はそれ以上意味を分割することができないのは、ある意味で当然である。この分割の仕方こそが、世界をどの断面で切るかを決定しており、一旦分割されたデータにいかなる処理を施してもデータを記述した枠組を越えた新たな視点による問題解決が行われることは原理的にあり得ない。つまり偏見に満ちたAIのシステムは自身の論理形式から逃れることはできないのである。このため、人間の助け無しには外界から新たな構造を持つ知識を得ることができない。つまり、AIの基本的コンセプトである世界の論理的記述が、そのままAIの致命的な限界に直結している。

AIの研究者はこの基本的困難にも関わらず、なんとかこの困難を乗り越えようとしているように見える。代表的な課題は具体的な例から一般的なルールを導出するような試みで、一般化能力に関わる。もちろんこの種の試みは限定された問題毎にはある程度成功するが、それは、その問題の解決によって導き出せるルールが記述しやすい論理的枠組を用いて問題を記述したから成功したと言える。日常の問題解決において、「問題をうまく表現できればその問題は半分解けている」と言われるが。これは、推論や思考において一般化や抽象化が適当に行われていればその問題は半分解けていると言うことである。つまり、AIはすでに人間が半分解いた問題の残りの半分を解いていると考えられる。一方、オーソドックスな論理記述の限界を打破するためにAIにファジーを取り込む様な試みもあるが、これも論理の許容範囲を広げることにより、記述できる問題の可能性を広げただけなので、これまで述べてきたAIの根本的な限界を乗り越えることはできない。

第2には、現在のAIは基本的に人間の意志に沿って何等かの処理を実行する受動的なシステムであり、自主的に何かテーマを見つけて問題解決したり、設計者の知らないところで思いもよらない計算を行う事はない。人間が様々なアイディアを生み出すために普段から色々考えているように、より人間に近い知能システムを作るためには、マシン自体が自発性を持つ必要がある。この能力を持つことによって、知能システムを自体が自発的に科学研究等を行うなどということもできるかもしれない。自発性をAIに導入することはおそらく可能だろう。本研究ではAIとは違った方法でこの能力を実現する方法を検討している。

第3には、AIには経験を通して形成される個性がないという欠点がある。人間や動物が面白いのは、それぞれ異なった意見を持っていて、個体によって得手不得手があるためである。今後、科学がいかに進歩しても、世界の全てを知ることはできないと思われるので、当然完全な知識を持った知能システムを作ることができない。例えば知能システムを裁判に使うことを考えてみよう。ここでは10台の知能システムがそれぞれに法律に基づいて判決をくだしその筋道を最終的な責任を持つ人間の裁判官に示すような、判決支援システムを考えることができる。このとき10台の知能システムが全く同じ様に考えたなら、10台ある意味がない。しかし、考え方は本来多様なものであり一通りの知能システムの判決例を参考にしすぎるのは危険すぎる。つまり、世の中の多くの問題は答えは一通りでないから、初期値だけでなく経験を通して様々な個性を作り出すことは重要である。この困難は、や

はりAIを改良することで克服されつつある。

以上従来のAI研究に対して批判を述べてきたが、研究者がこの限界を理解して適切な姿勢で研究を行なえば、その成果は有用なものと成り得る。つまり、今後新たな情報処理の方法が開発されたとしても、課題に対して適切な一般化でができるなら（ある論理的枠組の中で問題が解けると判っているなら）、おそらく論理的な手法を用いるのが最も効率的な方法であり、その座を明け渡す可能性は少ない。そして、この種の論理的処理が最も有効な分野が消滅することはないと予想される。

これまで説明してきたAIの限界は『一般化の問題』と考えられる。この問題はニューラルネットワークを使おうと使うまいと、知能システムを設計する上で避けられない重要な課題である。一般化が適切に行なわれているとは、現実のパターン空間から推論や判断に適したサブ空間への写像が適切に行なわれているものと考えられる。もし、そのサブ空間が行動に対応する出力にとって論理的に分割可能であれば、その後の処理はAIによって実行することができるし、さらにそのサブ空間が行動出力から見て線型分離であれば、単層のパーセプトロンでさえ判断ができる。ニューラルネットワークの研究における一般化の問題を考えるために、認識課題を例にとると、入力付近では意味付けがが明確だが、次第に高次になり連合と絡むに従って一般化の問題が色濃くなり、研究は混沌としてくる。

結局、これまで明確にされていなかった未開の研究領域がAI、ニューラルネットワーク、パターン処理の狭間に存在することを示唆する。言ってみればそれは『パターン推論』とでも言うべき研究対象であり、当然一般化を含む前段階とその後の段階がある。現在、私は一般化の問題に対する確かな答えを持ち合わせていないし、ニューラルネットワークが、この問題を解決する唯一最高の研究方法だとも考えていない。しかし、ニューラルネットワークは一般化能力を持つシステムの一例として存在し、研究の契機や道具として有用である。この視点から見ると、ニューラルネットワークは線型予測をベースとするパターン推論システムの一形態を実現する技術的手段と見なせる。

1-3. ニューラルネットワークとその特徴

ここで、本研究で中心的役割を果たすニューラルネットワークについて説明する。

1-3-1. ニューラルネットワークとその研究状況

ニューラルネットワークは生物の神経細胞を手本とし、基本的には非線形な入出力特性を持つユニットを多数結合したネットワークである。各ユニット内部の非線形性と線形和をとる部分との交互構造が特徴的である。初期の研究では、一種類のユニットにより構成された内部ループを持たない階層型のネットワークが対象とされたが、現在では内部ループを持つ相互結合型や、複数種類のユニットを持つネットワークの研究、更にはこれらいくつかのネットワークをモジュール化して全体のシステムとして高度の機能を実現しようとする試みなどと、その複雑さを増している。私の研究もほぼこの段階の研究であると言える。この種の複合的なシステムの設計を行うと各モジュールごとに要求される機能が異なる。すると、いくつかあるモジュールの記述のためには必ずしもニューラルネットワークが最適であるとは限らない。よって、私の研究においても都合によりニューラルネットワーク以外のモジュールを使用することがあるが、そのモジュールの機能が生物的に実現できることを要請している。また、この種のニューラルネットワークのカテゴリーには収まりきらない学際的な研究に対する興味が高まっているのも研究の現状である。

1-3-2. ニューラルネットワーク特有の長所

様々な人が自分の研究の立場から多くのことを述べているが、世間で言われているニューラルネットワークの代表的な特徴を六つほど揚げてみた。(合原,1988)

1) 学習や適応能力
   自己組織化能力。外部環境に合うように、自分自身を変化させて調整する能力。
2) パターン情報処理やあいまいなデータの処理
   これは、前記した一般化の問題に関わると思われる。
3) 連想能力
   ある入力パターンから別のパターンを想起するような能力。
4) フォールトトレラント性
   ある程度不確かな状況や誤りがある状況でもそれなりの対応ができる能力。
5) 並列性

処理を集中させず分散させた処理を行なうこと。

6)融通性

どんな状況でもまあまあの答を、短時間でなんとかひねりだそうとする能力。

これらがニューラルネットワーク特有の特徴であるかどうか考察する。

学習能力は、研究の初期の段階において生物のシナプスの可塑性に基づいて学習能力が導入されたというだけであって、本質的には処理過程におけるなんらかの変数を変化させることである。そして、この学習能力はAIや適応制御等でも用いられているので、ニューラルネットワーク特有の特徴であるとはいえない。

次のパターン情報を取り扱う能力については先に議論したので省略する。

連想能力はやはり必ずしもニューラルネットワークに頼る必要はない、これは一種の入出力関数を応用すれば実現できる。

フォールトトレラント性は2)の一般化能力に伴って実現できるだろう。

そして、並列化に関しては少なくとも、高速処理能力の実現できる可能性の探求と言う意義において議論の余地はない。

以上の考察より、上記6つの特徴の中で、あえてニューラルネットワークを使う理由となり得る本質的な特徴は一般化能力とと並列性だと思われる。並列性は理論的には問題なく実現されるが、現実はノイマン型コンピュータによるシミュレーターが主流であるから、現実的な内容は伴っていない。しかし、現在の状態でニューラルネットワークがハードウェア化されても人間のように振舞うというものでもないのでこの状況もやむを得ないが、将来的には並列化は実現できるだろう。むしろ、ニューラルネットワークをハードウェア化する意味があるといえるだけの基礎的な研究の積み重ねが必要であろう。そこで、おそらく一番有意義な能力が一般化能力である。しかし、この能力はまだ可能性として期待されている面が強く今後の理論応用両面での発展が望まれる。

本研究においてニューラルネットワークを意識した理由には、上記の意義以上にシステムを設計する際の思考ツールや足掛りとして大きな貢献をしている。つまり生物のシステムを見習うこと、それに伴って生物的な制限事項を想定することである。

1-3-3. ニューラルネットワークおよび関連の研究手法

ニューラルネットワークの研究.開発はその方法によって、(1) 脳の構造と機能の解明を目指す実験的研究、(2) 脳の計算原理の解明を目指す理論的研究、(3) 脳機能の工学的実現を目指すテクノロジーというように分類できる(Fig.1-2参照)。(合原,1988: 坂本,1983)

(1) 理論的研究

ミクロな研究は、神経細胞の生理的動作を説明するものと、神経回路モデルのの基本単位として、神経細胞の特性をできるだけ単純化したものがある。本論文に関係するのは後者のモデルでその研究はまず1940年から1960年代にかけて活発に行なわれた後、1982年頃から再び盛んになっている。モデルのユニットには、シナプスでの入力の空間的、時間的加重特性と、しきい値等の非線形特性が盛り込まれている。数年前のニューラルネットワークブームではホップフィールドのニューラルネットワーク、ボルツマンマシンおよび誤差逆伝播学習法がこの分野を賑わしたが、最近はこの種のモデルの理解が進み次第に実用化が進んでいると同時にその限界も見えてきたので新たなブレイクスルーが求められている。

マクロなアプローチとしては認知心理学と知識工学との境界領域である認知科学がこれに当たる。科学としての歴史が浅く、また人間の高度な知的活動を対象にするので他の分野からほとんど孤立していたが、近年神経回路を導入したモデルなども現れはじめ大変興味深い。

(2) 実験的研究

マクロな実験的研究としての、実験心理学では動物を対象として、スキナー箱のような理想的な環境を構成して学習過程を研究する。最近は、生化学、神経生理学と結び付き、学習による脳内物質の変化、学習によるニューロン活動の変化等のミクロなアプローチと一体になりつつある。心理物理実験では、主に人間を用いて、視覚、聴覚等の知覚（例えば錯視、文字認識、言語認識）を調べる。また、臨床病理学もこの一つだろう。

ミクロな研究は長い歴史がある。これまでの生理学、解剖学、組織学などの研究により脳の基本構成要素であるニューロンの動作、ニューラルネットワークの微細なアーキテクチャ、更には、知覚、運動、記憶、言語などの脳の高次機能に関するマクロな機能局在等多くの知見が蓄積されている。最近では、分子生物学、生化学、遺伝子工学などの更にミクロな手法も、ニューロンの動作のミクロなメカニズムの解析に用いられており、また、X線コンピュータ断層撮影法(X線CT)、核磁気共鳴映像(MRI)、ポジトロン断層撮影法(PET)、光学的計測、SQUID(Superconducting Quantum Interference Device)による磁気計測など、脳神経計測技術も大きな進展を見せている。

(3) テクノロジー

ニューラルネットワークを工学的に実現しようとする気運も活発であり、すでに一部ではビジネスの対象となっている。しかし、並列分散情報処理原理に基づく非ノイマン型コンピュータを目指すという、本来の勇ましいスローガンに反して、現在の商品としてのニューロコンピュータは、そのほとんどがノイマン型のカテゴリーに含まれるものであり、

将来の真のニューロコンピューターの研究開発のためのツールとして意義に留まっている。しかし、現状の基礎理論に基づいて専用のハードウェアを開発することに、あまり重きを置きすぎる必要は無いと思われる。

我々の研究手法は基本的に理論的なものであり、ミクロなレベルからマクロなレベルへの橋渡しをモジュール構造により実現している。

1-4. 本研究において配慮した点

本研究では入力から出力を含めた動作主体全体と環境により構成される大規模な系を取り扱う。そのため、シミュレーションでは適宜簡略化を行う必要がある。またミクロな生物学的知見や我々の日常の行動のあり方に沿ってモデルを構成しているので、事実を説明しようとする面と構成的な面が混在した研究内容と成っている。そこで、モデル化の各段階において、いかに本質を失わないようにするかが鍵となる。

いくつかの配慮した点は主に生物との対応から導出される。勿論、実用的なシステムを作る上では必ずしも生物と全く同じ構成をとる必要はないが、現在地上に存在する最も進化した知的システムが生物である以上、生物の脳を中心とする神経系の知見を考慮するのはある程度必要である。以下研究において考慮した点を述べる。

1-4-1. 情報伝播の制約

我々が興味を持つネットワークレベルの情報処理に関する研究では、構成ユニット間の情報伝播に対する制約が重要であると考えられる。なぜなら局所的な性質についてはニューラルネットワークの研究者の立場からは、その可能性を限定することができないのに対し、大域的な性質の予測は行うことが可能であるからである。しかも神経細胞に対する実験的研究の立場からは、局所的な情報処理に対する機能的な予測を促す成果は沢山あるが、制約事項を導き出せる知見は見あたらないので、ユニット内の局所的な情報処理能力は常識的な範囲で理想的に優れたものであると仮定する。すなわち、個々のユニット内部の情報伝播能力はユニット間の情報伝播能力に比べれば、遥かに大きいと考える。この条件は並列処理マシンを設計する工学的な視点から見た場合にも合理的である。

生物がその内部において情報を伝達する方法には様々なものがあるが、原理的には次の二つに分けられる。第１に生物が最も最初に採用した思われる、化学物質による情報伝播である。この方法は伝播速度が遅いので、ユニット内部や隣接ユニット間では伝達に関わる時間的遅れはある程度小さくて済むが、ネットワーク内の離れたユニット間では少なくとも数秒以上の大きな時間遅れが伴う。しかし情報量は化学物質の多様性によりある程度大きくすることができる。そして、ある化学物質に対する感受性を変化させることで、隣接していないユニット間でも情報伝達相手の特定を実現することができる。つまり、バスラインを用いた通信のように送信側と受信側のアドレスがユニークに設定されていれば任意のユニット間で通信を行うことができる。しかし、この方法は送信側や受信側の性質が同じだと区別できない。実際の脳の中で全てのユニットにユニークなアドレスを割り振る

とは思えない。せいぜい性質の異なったユニット群に対して異なった情報を送る程度であろう。第２の方法は電気信号による。動物に見られる全体的な動作を行うためにはどうしても距離の離れたユニットの共調的な活動が必要である。この問題を解決するために分化したのが神経細胞であり、ここでの新発明は、空間的に広がったユニットの電気パルスによる情報伝播である。これにより、距離の離れたユニットを実質上近接させることができた。この方法は伝播速度の点で有利である反面、伝播過程で使用される電気パルスが一次元的な量であるため伝達できる情報量は少ない。たとえ、時間的な相関などを利用することができたとしてもそれほど大きな情報量を担うことはできない。伝達相手の特定は、シナプス結合によって調整できるが、情報量が少ないので細かい制御を行うことはできず、その効果を十分に発揮することはできない。以上の考察をまとめると、Table 1-1のようになる。

	化学物質信号		電気信号
	隣接	非隣接	
情報量	多い	やや多い	少ない
伝播遅延	小さい	大きい	小さい
特定性	できる	できない(*)	ややできる

(*) ユニットの種類毎にはできる

Table 1-1 生体内部の情報伝播特性

これらの要請をまとめると、局所的な範囲ではお互いに必要な情報をいくらでも交換できるが、大域的な情報交換は低速で大量か、高速で小量のどちらかを選ばざるを得ない。

よって、ある程度大きなネットワークの動作状態や学習を一つのユニットやスーパバイザーにより完全に制御することはできないし、一つのユニットが全ての情報を集めることができないので、完全なスーパーバイザーとなることもできない。

1-4-2. 環境と動作主体の相互フィードバック

検討すべき系は、環境からの刺激を入力信号として受け入れ、外部環境に対して出力信号を送り出す動作主体と、動作主体の出力と以前の自身の内部状態に依存して状態を遷移させ、それに対応した出力を動作主体に送り出す環境とにより構成される。

この系ではあからさまな教師やスーパーバイザーは存在せず動作主体が持っている評価

関数に従って、自己組織化を進める。この問題設定は、岡部研究室の伝統的な出発点でありこれまで多くの研究が為されている。

1-4-3. 遅れを持つ批判信号

前記した環境と動作主体のみによる系では環境中には教師が存在しないと考えるべきなので、評価をするのも動作主体自身と考えざるを得ない。だからといって、動作主体の持っている評価基準は環境と全く関係が無いわけではなく、むしろ生物の進化における自然淘汰の過程を通して深い関係を保っている。つまり、適切な評価基準を発現する遺伝子を持った個体が成功し生き残るのである。(ﾄﾞｰｷﾝｽ,1989)

生物の生活を考えた場合、その内部の評価基準が一般のバックプロパゲーションの場合のように、答える毎に正しい解答を教えてくれる教師であるとは思えない。遺伝子にあらかじめ記述されている評価は食欲や性欲など基本的な価値のみを知っているだけであろうから、どの様に行動すれば成功するかを教えることはせず、ただ文句だけを言う批判信号(critic)であろう。また、この批判信号はあらゆる出力に対して評価を与えることはできないので、その信号は常に出力されるわけではない。よって、ある動作に対する批判信号は時間的遅れを持つ。

批判信号を用いた先駆的な研究はWidrow, Klopf, Barto, Suttonなどによって為されているが(Barto et al, 1981: Barto et al, 1983: Klopf, 1986, 1987, 1988: Muro, 1986: Sutton et al, 1981: Werbos, 1987: Widrow et al,1973)、これらの、批判信号が一般に動作主体の外から与えられていることからスーパーバイザーの必要があり、生物のシステムとは異なると主張する向きもある。確かにこれはある意味でスーパーバイザーであるが、先に述べたような理由でこれが生物的でないと考えるのは適切ではない。批判信号は教師信号と比べて遥かに簡単であるから、Genetic algolithmのような原理で環境から動作主体に刷り込まれ、今では動作主体は生まれながらに適当な批判信号源を持っていても何の不思議もない。

1-4-4. 逐次的な学習

これまでの問題設定からも明らかな様に、学習は動作主体が環境と相互作用する過程を通して行なわれる。つまり、動作主体は経験を通して次第に適応的な行動ができるようになる。よって、ニューラルネットワークに対する学習はある時点で多数の入出力関係を覚

え込ませたり、同じ入出力関係を計画的に何度も提示したりするわけではなく、あくまで環境と関係を持ちながら自然に学習を行なう方法をとる。

1-5. モデレーショニズムから見た本研究の位置付け

これまで本研究室ではモデレーショニズムというテーマで多くのニューラルネットワークの研究が行われた、この考えは岡部により提案されたもので、その原理は『生物は適度な刺激を好む』で、個々の機能素子による分散学習制御を実現する可能性を持つ画期的な原理である。この原理は1980年代初頭に岡部、坂本(1983)、北川らによる心理物理実験により、ある実験条件では成り立つことが示された。その後、深谷、大輪らにより理論的研究が進められている(Fukaya,1986: Okabe,1986,1988: 深谷,1988: 大輪,1989)。深谷の研究では環境が時間的な意味での教師信号を送っていると考えられる点で、また大輪の研究では陽に環境中に教師を仮定している点で、初期の思想が充分に生かされていない。また最近の山根、甲原らの研究によれば反射弓程度の簡単なシステムではある程度成功することがシミュレーションにより確かめられたが、より高度なシステムやより一般的な課題に対してどの様に適用して行くかは大きな課題として残されている。また、この方法論の欠点はユニットレベルの動作原理から行動レベルの結果を引き出すので、構造的に中間的なレベルにおける意味付けができない。そのため、研究を段階的に行うことができず、シミュレーションを行ってもどこまでが成功しているのかを判定できないのが泣き所となっているように思われる。また、この欠点のために実験的手法により研究を行っている研究者に役立つ情報を生み出せない状況を作り出している。さらに、動作主体をホモ（単一種類）なユニットによって構成する方法には限界があるのではないかという見解もしばしば耳にする。なお、各ユニットが独自の評価基準を持つという考え方はKlopf,Barto(1985)らによっても検討されているが、完全に学習過程を独立分散させた手法は他に例を見ない。

モデレーショニズムと本研究の基本的な出発点は、先に述べた研究における留意点のうち、(1)生物的に実現可能な構成、(2)環境と動作主体の相互フィードバック、(4)逐次的な学習、の３点つについては共通している。

これに対してシステムを統括する評価関数については、異なった立場をとっている。モデレーショニズムでは評価関数はネットワーク内の各ユニットに独立に存在するのに対して、本研究における(3)遅れを持つ批判信号では動作主体全体にとって唯一の評価関数である。生物進化の歴史に目を向けてみれば、単純な段階では確かに個々の細胞が独自の評価に従って行動しているが、多細胞生物では明かに各細胞は自分自身の生存確率を増加させる利己的な評価基準に従っているとはいえない。よって、動作主体内に共通な合目的評価基準が存在し得る。

つまり、本研究は基本的には従来の本研究室の基本的戦略を踏襲しつつ、モデレーショニズムのかかえる問題のいくつかを避けることができる方法となっている。

1-6. 研究の方針

以後、本論文では主に次の三つの章に分けて説明を行う。

第２章　価値観を持つ知能システム

はじめに自発的に思考するマシンを設計するための世界観についての概念的な枠組をどう与えるべきかを考察し、その後に仮想的な環境内で動作主体がとる行動から、動作主体に要求される能力とそれを実現する基本的な構造を思考シミュレーションによって導き出した。動作主体にはおもに認識と思考、価値観、行動に関係の深い三つの基本的モジュールが必要であることが示された。さらに、自発性を実現する為に追加すべき構造も検討された。

第３章　ネットワークの状態コントロール

研究において考慮した点の(1)情報伝播の制約、での議論を基に、ニューラルネットワークの大域的制御で何が可能なのかを調べるた。この研究により得られた結果は学習による効果を全く利用しなくても、ユニットのダイナミクスの調整と補助的な抑制層の付加によりネットワークの活動状態が制御できることであり。最も典型的な効果としてはネットワーク内で発火しているユニットの数を制御するような場合である。

第４章　迷路のようなシミュレーション

価値観を自己生成する機構のシミュレーションを試みた。動作主体を簡略化しつつも、環境を含む系全体を取り扱うことによりその有効性を確認した。

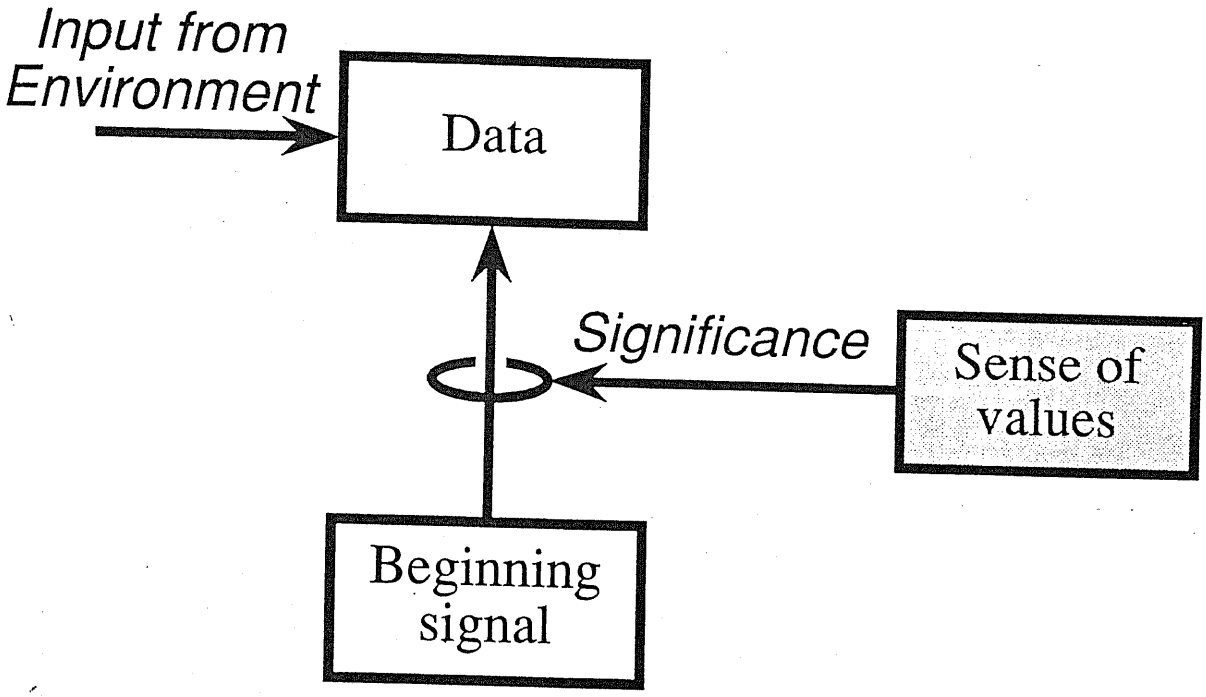


Fig. 1-1 自発的に思考を行なうシステム

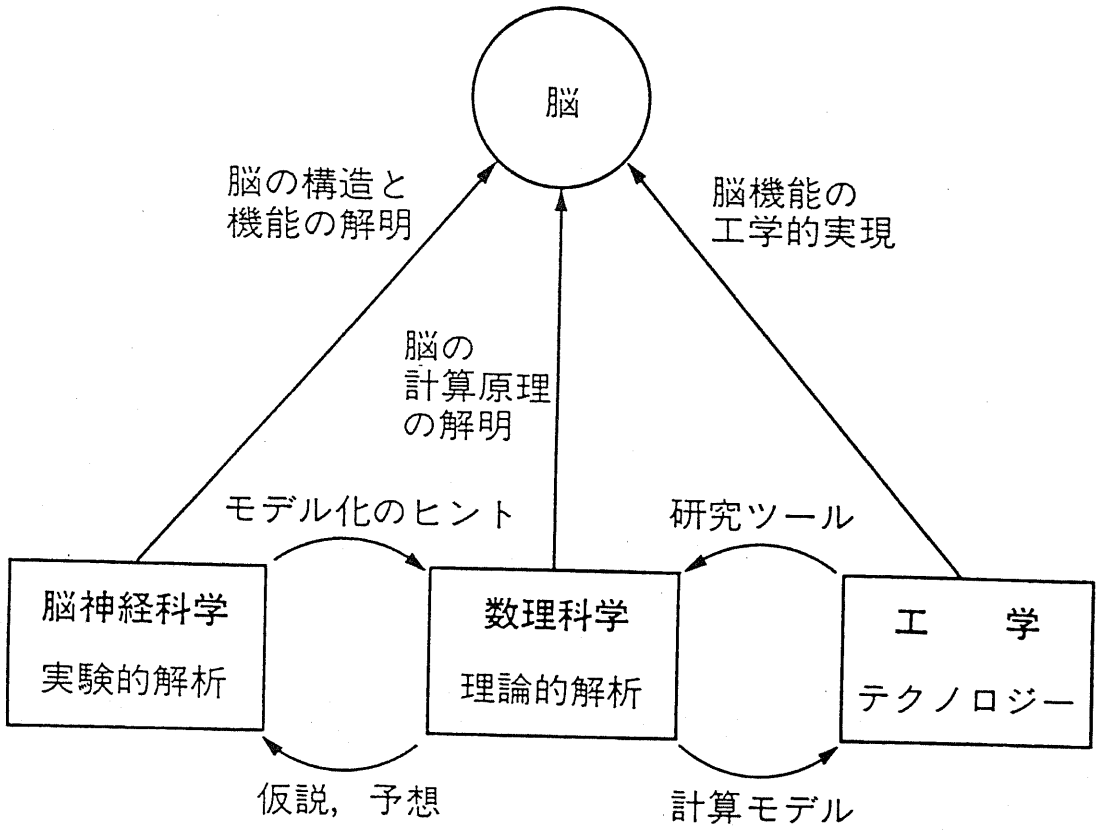


Fig. 1-2 ニューラルネットワーク研究マップ

第２章

価値観を持つ知能システム

2-1.環境と知識そして思考と意識

本節では、本論文における思考の捕らえ方を明確にするために、思考の対象となる知識および環境との関係、さらには思考に深く関わる意識について考察した。

2-1-1. 知識

第１章で説明したように現実の世界はパターン情報により表現されているので、知識もパターン情報に基づくことになる。パターン情報の様子を言語で表現するのは無理があるが、例えばFig.2-1に示すように、リンゴは赤、球形等と関連づけられ、赤は苺、血液等と関連づけられている。リンゴを主体だと思えば赤が属性で、赤が主体だと思えばリンゴが属性である。つまりパターン的な認識では連想的な関連付けは果てがなく、本来的にどちらが主でどちらが従という事はない。ただ相互の関係のみが本質的な意味を持ち、なんら枠組を持たない。

次に知識を構成するのに役立てることができる情報は明らかにシステムの過去の経験のみである。ところが、知識はあらゆる環境のもとで存在するかと問うと、入力パターンが完全にランダムな環境ではシステムがいかなる戦略をとっても評価関数を増加させることはできない事から、明らかに知識は存在しない。つまり知識は環境に依存するのである。実際良く考察してみれば知識が有効となる必須の条件は『出力と入力に因果的相関が存在する』ということである。さらに、Fig.2-2に示すように、過去から現在に到る入出力間の全ての相関、入力内の及び出力内の相関が知識となり得る。つまり入出力パターンに相関が存在することが知識が存在するための条件である。さらに知識の信頼性は出現頻度と普遍性などに依存する。

以上の様な相関は最も基本的な知識ではあるが、我々の知識は単なる出来事の記憶のみ

ではなく、より一般的な抽象度の高い概念を含んでいる。つまり、我々は抽象化能力を利用して現実に体験する具体的なパターンから一般性の高い共通の相関を抽出することにより抽象的な概念を導出することができる。Fig.2-3に示すように、抽象化が進むほどその概念に含まれる共通属性の数は減少すると期待され、高度な抽象化は言語さらには数学などの思考法に到達する。適応においての抽象化の第１の利点は予測能力の向上である。これにより過去の経験をより柔軟に新たな問題に活用することができる。実際我々はあらゆる場面で少なからず抽象化を行っている。言語を用いて何かを説明することを考えてみると、言語は抽象的なシンボルであるのに対して世界はパターンによって記述されるので、何かを言葉で言い表すことは、本質的に分解不可能な世界を特定の方法で分割したシンボルと言う名の断片で再構成することである。人間の言葉による説明がしばしば多くの誤解を招いたりするのは、このためである。シンボルによる説明は決して完全にはなり得ないのである。

結局、知識とは動作主体が環境との相互作用から得たパターン情報内の相関およびそれに対して抽象化を行った情報である。

2-1-2. 思考（知識処理）

思考とは何であろうか。典型的にはパターン情報として貯えられた知識を未知の状況に適用する操作であり、問題解決のために適切な抽象化を用いて推論を行う。そこで、思考を定義する上で二つの重要なポイントを指摘する。一つは思考が原理的に内部に貯えられた情報のみを利用して機能できる点である。このことから、思考は外界との相互作用により予めネットワーク内に貯えられた情報に対して行われる操作と考えざるを得ない。二つ目のポイントは前説で述べたパターン情報の特性に負うもので、情報の本質が相関のみという点である。以上二つのポイントから思考は何等かの形で相関関係を強調する操作であると推察できる。そこで、我々は

『思考とは相関の強調である。』

と仮定した。

この強調により情報はより有効に利用できると考えられる。例えば、Fig.2-4にを用いて説明すると「山でリンゴを食べたら美味しかった」(α)というパターンにより表される記憶と、「海でリンゴを食べたら美味しかった」(β)という記憶が、漠然と存在しても、空腹時に何をすべきかわからない。しかし、二つの事実から引き出される「リンゴは美味しい」という、より一般的な関係を知っていればをリンゴを目的とすべきであることを判る

のは容易である。

この知識を導出するには思考に相当する相関の強調を行なえばよい。そのためには、たとえばFig.2-4のように、美味しいという概念を軸（2-1-3.参照）として、関連する山や海での体験(α),(β)を繰り返し再現することで、美味しいとリンゴの間の相関を強化することができる。

つまり、思考はある概念を軸として経験により貯えられた情報を連想的かつ連続的に再現することにより、情報自身の相関の強調を行うことである。つまり、内部に貯えられた情報を利用しやすい形に再編成する操作である。なお以上説明した例は静的な経験((α),(β))についてであるが、短時間に動的に移り代るパターンも相関を持つので、同様の方法で思考を行うことができる。

この思考により期待される典型的な効果は、Fig.2-5に示す相関の結合として現れるだろう。前記のように「リンゴが美味しい」(A→B)という知識を獲得した上に、それ以前に、「リンゴはお金で買える」(B→C)という知識を持っていたとする。すると思考によって2つの知識と結合が起こり、「美味しい思いをするにはお金が必要」(A→C)という考えに及ぶ。ニューラルネットワークではこの例の様な相関の結合は自然にも起こりえるが。思考はより適切にこの機能を実現する。他の実例としては、数学の問題を解く場合に様々な公式を結び付けようと考えることや標準的な三段論法など多くの思考がこの描像で表現できる。

2-1-3. 意識

2-1-2.で思考を行なう際に注意を固定するための軸が必要であることを述べた。この作用は思考しているネットワークに対して想起し易いパターンを制限することで、これこそが意識の作用に対応するものである。注意を与えるべきネットワーク上の目標となる状態は、システムの目標と関連の深い状態である。そこで、思考すべき課題であるかを示す重要性を何等かの形で与える必要がある。しかしこの基準は外から与えてやるのではなくシステム自身が知る必要がある。そこで重要性を記述する価値機構を2-3.で検討する。

2-2. 環境に対する適応の戦略

知識や思考は動作と結びついてはじめて意味を為すので、行動まで含めた環境と相互作用するモデルを検討する。

適応的なシステムの最も単純な記述はFIg.2-6中の入出力回路（出力ニューラルネットワーク: ONN）のみで記述できる。我々の研究ではこれに二つのモジュールを付加した。一つは外界の相関を保持することで環境に対するモデル化を行い、さらに思考を行う場を提供する認識連合モジュール(CAM)。もう一つはパターンに対する重要性を含む価値体系を構築する価値評価モジュール(VAM)である。

2-2-1. 出力ニューラルネットワーク(ONN)

本モジュールは外界からの直接の入力及び認識連合モジュールの出力の何れも取り込むことができる。出力部に教師信号は与えられないので、その学習は入出力間の相関に依存し、価値評価モジュールからの評価信号による強化学習が行われる。つまり、報酬性の増加および嫌悪性の減少時には行動の強化、報酬性の減少および嫌悪性の増加時には行動の抑制が行われる。

2-2-2. 認識連合モジュール(Cognition and association module : CAM)

本モジュールは外界からの直接の入力を取り込み、その相関を保持する相互結合型のニューラルネットワークを基本とし、世界モデルをその内部に構築することを目的としている。この際に出力ニューラルネットワークの出力を入力に含めることは当然可能である。ここでは入力の相関を記憶するので学習法則は基本的に相関学習であるが、強化学習と同様に他のシグナルによってその学習係数が制御され、教師信号は存在しない。この学習係数の制御の必要な理由は、ネットワークの能力は限りがあるのであらゆる相関を記憶することはできないから、そこで入力された情報の有意義な部分に対して集中的モデル化を行う必要がある。この選択の為には主に２つの基準が思い当たる。一つはシステムに対する重要度の基準、二つ目は新規性の基準である。一つ目の基準については価値評価モジュールからの重要度信号により学習の進行を制御する。二つ目についてはまだ検討が不充分であるが、認識連合モジュール自身の活動状態を監視することで感知することができるので

はないかと考えている。

2-2-2-1. 認識と連合

このモジュールでは認識と連合両方の機能を合わせてあるが、これはともに環境から取り入れた生の情報を行動決定に結び付け易く変換する作業である点で共通である。方法論としては、(1) 先天的な情報や環境の統計的性質を利用して特徴抽出的な自己組織化を行う方法(福島, 1989: ﾘﾝｽｷｰ, 1989)と、(2) 出力の情報を利用し環境の多用性に応じた柔軟な自己組織化を行う方法、がある。前者は基本的な意味での認識であり一般に教師無し学習(Unsupervised learning)として議論されている。後者はむしろ連合にとしての機能である。生体では二つの機能が混在しながら次第に移り代る可能性が高い。

2-2-2-2. 認識連合モジュールにおける漸進的モデリング

CAMはつまり、生の入力空間を取り扱いやすいサブ空間に変換する機能を請け負うわけだが、CAMは始めから完成しているのではなく、経験を通して高度な認識と世界観を獲得する。一方ONNは未完成なCAMの信号を用いてそれなりに行動を行う必要がある。よって、認識連合モジュールは完成後に機能を果たすのではなく、ONNに利用されつつ漸進的に自己組織化を進めると同時に、学習途上の不完全な段階でもそれなりに動作する能力が求められる。

この際モジュールをニューラルネットワークとして具体化すると、従来からのバックプロパゲーションの手法をそのまま用いると不都合が生じることをFig.2-7を用いて説明する。例えばユニットＡから出力ニューラルネットワークへの結合によってある行動を行う関係がすでに完成した後で、ユニットＢからの結合によって行動の制御を調節するためにユニットＢの反応性を変更したくなったとしよう、この際ユニットＡの入力結合も変更されてしまうのでその反応性も変化してしまう。すると既に完成し成功していた古い入出力関係に悪い影響を与えてしまう可能性が大きい。さらにもう一つのCAM内部の問題を指摘しておく。つまり、学習を始めてから有る程度時間が経過した後では、すでに経験に対応する内部モデルが作られると期待される。一方、新しいより複雑な処理は古い低次の処理の結果を利用して行われるため、低次の処理がその時点で変化すると、それを用いる上位の処理は大きな影響を受ける。よって、実際的な学習では入力に近い既に他の部分やモジュールに利用されている低次の部分はよほどの矛盾が生じない限り変更せず、未だ出力モジュールと明確な結合を持っていないそれ以外の部分の可塑性を大きくする漸進的なモデリングの手法が必要とされるだろう。

2-3. 価値評価モジュール(Value assessing module : VAM)

2-3-1. 価値評価モジュールの設計

2-1-3.で重要さを表現する価値が思考に必要であることを示した。これらの価値観を形作るのが価値評価モジュールである。本研究では価値観は報酬性と嫌悪性の２種類の性質の関係によって成り立つと考える(小野,1989,1990)。そしてその基となるのは第１章で説明した時間遅れを持つ批判信号であると考えた。ここで、当然批判信号にも報酬性と嫌悪性がある。

2-3-1-1. 基本モデルと基本価値評価モジュール

Fig.2-8の基本モデル(Basic model)の範囲では、基本価値評価モジュール(PVAM)は価値評価モジュールそのものである(中野,1987: 佐賀,1983)。基本価値評価モジュールはシステム内に備えられた批判信号源で動物における快感中枢などに対応する。価値評価モジュールの原型であるこのモデルでも、既にFig.2-9に示すように入力に応じ報酬性と嫌悪性に対応する２値ベクトルを出力する。Fig.2-9の変化感知モジュール(Variation sense detective modeule : VSDM)内における報酬増加感知器(Increase reward sense detector : IRS)と嫌悪減少感知器(Decrease punishment sense detector : DPS)は、それぞれ報酬性信号の増加と嫌悪性信号の減少に反応する。これらの出力は出力ニューラルネットワーク(ONN)の学習を制御することになる。ここで、出力ニューラルネットワークに付属している学習促進器(Learning prompter : LP)はそこへの入力が学習促進の刺激として利用されることを示している。

詳しい説明と思考の助けのために思考シミュレーション用の仮想的な環境を設定した。このイメージはFig.2-10の様でその関係は　Fig.2-11に示された。例えばりんご(Apple)の刺激があるときに食べる(eat)を出力すれば入力刺激は栄養物(Food)に変化するのである。

ここではFig.2-12が示すように、栄養物(Food)と痛み(Pain)がそれぞれ、報酬性と嫌悪性の基本的価値に対応し、餌を食べたときや痛みを感じた時に、原始価値評価モジュールが対応する出力を行う。

基本モデルがこの環境中で、さまよい歩いた場合の動作について考えてみる。まず、はじめてリンゴを食べた場合、基本価値評価モジュールが報酬信号を出力する。この時、出力ニューラルネットワークは報酬性価値の増加に促されて、現在または少し過去の相関関係を強化し、リンゴを見たら食べるという動作を覚える。さらに、認識連合モジュールでも報酬性価値つまり重要性の価値が増大した状態となるのでリンゴのパターンを覚える。しかし、その後リンゴの木からリンゴを採るという動作を行っても、基本価値評価モジュ

ールが反応しないため、その動作やリンゴの木に対応するパターンは覚えられない。嫌悪性についてはすこし異なっていて、始めてクモに触って痛い思いをしたときに価値評価モジュールから嫌悪性信号が出力されるので、クモのパターンを覚えることはできるが、適切な動作を覚えることはできない。

結局、基本モデルでは報酬性については基本的価値と強く結びついた極めて限定された適応のみが可能である。一方嫌悪性については行動を学習できず、実質的に全く機能しない。

2-3-1-2.二次価値概念の導入

上記の問題を避けるめに基本価値から派生する二次価値の概念を導入する。先に述べたように情報の中で本質的な意味を担うのは相関であるから価値を与えるべき対象も相関に基づく。

Fig.2-8をもう一度見ると、基本モデルでの限界はリンゴを見たときに価値評価モジュールが出力を行わなかった点にある。そこでこの問題を克服するために、一次モデル(First order model)では、二次価値評価モジュール(Secondary value assessing module : SVAM)を付加した。二次価値評価モジュールの学習制御器(LC)に入力がある場合には、モジュールは現在又は少し前の時刻における入力刺激に対して反応するように学習を行う。すると、リンゴの入力の後に栄養物刺激があると基本価値モジュールの出力が二次価値モジュールの学習を促進し、二次価値モジュールは直前に入力されたリンゴに対して反応することを憶える。つまり、二次価値は基本価値と相関の高い入力パターンに対して付加される。全体図はFig. 2-13に示された。二次価値モジュールは報酬性と嫌悪性のそれぞれに対する新たな価値を覚えるために二次報酬発生器(Secondary reward generator : SRG)と二次嫌悪発生器(Secondary punishment generator : SPG)を備えた。反学習促進器(Reverse learning prompter : RLP)は通常とは逆向きに学習を行わせる。

一次価値モデルの仮想環境中での振舞いは始めてリンゴを見つけるまではほとんど基本モデルと同じである。つまり、リンゴを見たら食べるという動作を覚えた上に、リンゴのパターンを記憶する。ただしこの際、以前のモデルと異なるのは二次価値モジュールがFig. 2-14に示すようにリンゴに反応することを覚える点である。つまり、リンゴに報酬性の価値が与えられるのである。このため次の機会にリンゴをリンゴの木から採った時に報酬性信号が価値評価モジュールから出力されるので、リンゴの木を見たらリンゴを採るという動作も学習すると同時に、リンゴの木のパターンも記憶される。しかしこのモデルでもこれ以上複雑なリンゴ村に行くような動作を覚えることはできない。一方嫌悪性についてはクモに対して嫌悪性の価値を与えることができるので、クモを見たときに偶然避ける(avoid)ことができたならば、嫌悪性信号が減少するのでその変化が嫌悪減少感知器により

捕らえられ、クモを見たら避けるという動作が学習される。しかし、状況によってはさらに早い段階から危険を回避した方が良い場合もあるが、このモデルではその様な状況には対応できない。

結局このモデルでは報酬性に関しても嫌悪性に関しても性能が改善されたが、より複雑な行動を学習するには不十分である。

そこで、Fig.2-8の高次モデル(Higher order model)に示すように二次価値評価モジュールのオーダーを増やして高次の価値体系を獲得することができる。しかしこの方法には、二つの問題がある。一つには価値のオーダーを増やすのにしたがって二次価値評価モジュールの数を増やす必要があるる。二つ目には同じ入力に対して複数個のモジュールに記憶されることがあるので記憶効率が良くない点である。

2-3-1-3. 再帰的な構造

前節で述べた問題点を回避しつつ、高次の二次価値を実現するために、二次価値モジュールの出力を自身の学習制御器にフィードバックする再帰的な構造を提案する（Fig. 2-15参照）。この関係を報酬性と嫌悪性の二つに分けて整理したのがFig. 2-16である。入力はシステムに対する生の入力とさらには認識連合モジュールからの出力を含むベクトル信号である。一方出力は報酬性と嫌悪性に対応する二次元ベクトルとなる。報酬性の信号は二次報酬発生器への入力結合の学習を制御する学習促進器に入力されると同時に二次嫌悪発生器への入力結合の学習を制御する反学習促進器に入力される。嫌悪性の信号に関してはこれと全く逆の状態である。そして、システム全体をFig.2-17に表した。

再帰モデルはFig.2-18に示すように、一次モデルと同様の動作を行える上に、リンゴの木からリンゴを採った際にリンゴの木自体に報酬性の価値を与えるので、次にリンゴ村に行きリンゴの木を見たときに、リンゴ村に行くこととリンゴ村のパターンを覚え、さらにリンゴ村に報酬性の価値を与える。

報酬性に関するこの様な連鎖は果てしなく拡張できるので、このモデルは非常に多彩な行動及び思考を行う可能性を持つ。一方、嫌悪性に関してもより早い段階から危険を回避する能力を実現することができるが、この能力を拡張し過ぎると必要以上に慎重になり可能な行動の選択の巾を必要以上に狭めてしまう可能性を持つ。

2-3-2. 他の研究者によるモデルとの比較

我々の提案したモデルは、出力ニューラルネットワークが、基本的に入出力の相関に依

存した学習を行うという点でいくつかの研究(Barto et al, 1983: Jameson, 1990)と関係が深い。また外界に対するモデルを作るという発想においては他のいくつかの研究(Muro, 1987: Werbos, 1987, 1989: Shmidhuber, 1990A, 1990B)と共通点を持つ。また、主に Barto, Sutton, Werbos 等により、強化信号(Reinforce signal or Critic signal)の予測という立場から導出された学習則は、今回我々が価値観の観点から得た再帰的構成との関係が興味深い。

Fig.2-19に示す Werbos のモデルは、3個のモジュールを持つという点で我々のモデルとの比較において都合がよい。このモデルの目標は、評価Uの時間積分を最小または最大化する事であり、その予測として二次評価Jが導出され、それを基にエミュレータ・ネットワークによりコントロール信号に対する教師信号を生成している。

このモデルと我々のモデルの主な相違点を考える。まず世界モデルを取り扱うエミュレータ・ネットワークについては、この方法ではコントロール信号つまり行動を徐々に変更することになる。しかし、人間の採っている方法がこの様な形であるとは思い難い、なぜなら我々の行動の選び方はもっととびとびで、同一の状況に対してでさえかなり異なった動作を試す。この様な機能を実現するにはむしろ、我々のモデルにおける認識連合モジュールのように単に入力の相関を記憶しておいて2-1.で説明した思考の操作を組み合わせた方が実現の可能性が高いのではないかと思われる。

またWerbosらのモデルでは従来の多くのシステムと同様に全ての状況においてに対して同じ量の処理を行う。直感的には色々な状況に対して同じ正確さで処理してしまうと同時に、何を考える場合でもあらゆることを考慮してしまう。このような方法は、系が複雑になると計算量が爆発的に増大する問題を持つ。一方、人間は使用可能な時間に応じて、できるだけ少ない情報を用いて処理を行おうとする。つまり、時間をかけるべき処理は、動作主体にとって重要な意味を持つと同時に判断の難しい、報酬性と嫌悪性を同時に含む課題である。ところが標準的な価値関数では報酬性と嫌悪性がともに大きくても、打ち消しあうために無価値だと見なされてしまう問題がある。そこで、通常の評価関数の他に、何に対してより多くの処理時間をかけるべきか、規定する重要性を表す価値がシステム内に記述されている。例えば生死に関わる重要な判断と、歯磨きと洗顔のどちらを先にすべきかの判断を同じ程度まじめに考えるのは無意味である。この点に関してつまり、通常の評価関数と併せて重要性を表す価値が必要である。

そこで、我々は次に価値評価モジュールによって与えられる価値観の中に重要性を表す次元があることを示す。

2-3-3. 価値観の次元

価値観の次元はその構成の起源に基づいて、報酬性と嫌悪性の二次元ベクトルによって

記述される。一方、これを利用する点からはまず通常の評価関数として報酬性と嫌悪性の差に相当する量が利用される。一方前節で必要性を指摘した重要性は、報酬性と嫌悪性の何れが増大しても増大すると思われる。例えば危険な山の向こうに沢山の黄金があるならば人々はいかにしてその山を越えようかと苦慮するだろう、これはつまり報酬と嫌悪の両方が大きい場合にあたる。逆に普段いつでも簡単にてにできる空気の様なものに対しては報酬性も嫌悪性もほとんど無く重要性は認められない。

この関係は直感的にFig. 2-20の様に表されるので、評価関数と重要性の２値ベクトルは報酬性と嫌悪性の２値ベクトルとして変換されその関係は最も単純化した場合

評価 = 報酬性 - 嫌悪性

重要性 = 報酬性 + 嫌悪性

となるだろう。

つまり我々の用いた報酬性と嫌悪性の２値ベクトルの表現は重要性の価値を含んでいる。

なお、これとは別にある処理においてどの情報を関連づけて取り扱うかを決定する重要度もあるが、このシステムでは認識連合モジュール内の世界モデルによって記述されるものとしている。この例としては数学の問題を考えるときには、関係のない今夜の挽御飯のメニューや夏休みの旅行の予定などは重要度が低く、数学の公式は重要度が高いこの様な場合に関係のないものを同時に考えるのは混乱を招くだけである。

2-4. 自発的に思考するシステム

2-4-1. 思考の方法

2-1-2.で説明した思考のプロセスをこのモデルで実現するために、認識連合モジュールと価値評価モジュールの相互作用を利用する方法を提案する。思考における意識の機能を説明するためにFig. 2-21においては認識連合モジュールと価値評価モジュールの一部を表した。二次価値評価モジュールは、複数のユニットにより構成される階層型のネットワークであり、その中のユニットの反応性はモジュールへの入力パターン毎に異なる。従来どおり価値評価モジュールの出力は出力ニューラルネットワークの学習制御に用いられる。この場合、価値評価モジュールの報酬性と嫌悪性のそれぞれのユニットの活動の総和の差が強化信号として利用される。

一方新たに価値評価モジュールの出力自体を認識連合モジュールにフィードバックする結合を追加した。この機能は2-1.で述べたように認識連合モジュールで再現されるパターンを、注意の軸と関連の深いパターンに制御するためのものである。認識連合モジュール内の再現し易いパターンを価値評価モジュールの出力により制御する様子をFig. 2-22に示す。ここでは認識連合モジュール中の状態空間とその現在の状態及び状態に働く力をポテンシャルの形で表現している。実際には認識連合モジュール内の結合が状態空間の状態遷移ベクトル場に反映されるのでこの形状は複雑になるが、ここではそれらの効果は無視している。

表示されたポテンシャルは本質的に二種類に分離される。一つは認識連合モジュールが連想を行なうことが可能なように、素早い状態を変化促す斥力的な自己ポテンシャルである。自発的な連想を起こすために状態を不安定化する方法としては、この様に自己抑制ポテンシャルを用いる方法のほかにも、確率的遷移、シナプスやユニットの疲労等を用いてもよい。

もう一つは価値評価モジュールの出力により生成される引力として作用するスカラーポテンシャルである。これは思考の軸（意識）として作用するので、あまり速く変化し過ぎては不都合である。しかし、思考においては連想すべきパターンが単調な繰り返しにならないように意識の注意点を関連する対象に徐々に変化させる必要がある。よって、このポテンシャルもゆっくりと変化させる。

二つのモジュール間に存在する学習則の一方は、認識連合モジュールから価値評価モジュールへの価値に依存した相関的な学習であった。そこで、価値評価モジュールから認識連合モジュールへの学習も相関的に行うこととする。すると認識連合モジュール上の価値のあるパターンにポテンシャルが形成することができる。これにより価値と世界モデルとの相互的な連想関係を成立させることができる。

さて、以前に述べた様に効率的な適応を行うためには重要な処理に適切な時間に割り当てる必要がある。目の前で重大な事件が起これば対処すべき問題はその事件であるから注意をその事件に向けるべきであり、眼前に直ちに対処すべき事件がない場合には過去における未解決の課題を処理すべきである。

この動作をFig. 2-23を用いて説明する。認識連合モジュールの状態空間中において価値評価モジュールからの結合により空間のある部分（重要性をもつ部分）は潜在的に引力ポテンシャルを発生することができるが、そのほかの部分はこの潜在的なポテンシャルを持たない。そこで、もし外部からの入力にシステムにとって重要な情報が含まれていれば価値評価モジュールの出力パターンは大きく影響を受け、認識連合モジュールの状態は新たにその入力に関連する状態にトラップされ、つまりは目の前の事件の注意を集中する。逆に外部からの入力に重要な価値が含まれていなければ、その行き先にトラップされることはないので、眼前の事実ではなくこれまでの認識連合モジュールの出力が含む重要性にトラップされたままで、その後は近傍で次第に状態を変化させる。この後半の動作が思考にあたる動作である。なお、思考するモデルの全体図は、Fig.2-24に示した。

このモデルに従えば、意識は各瞬間における注意点であり、認識連合モジュールと価値評価モジュールの出力パターンの相互的なフィードバックにより発生する共鳴子である。また、思考は認識連合モジュールに形成された世界像と価値評価モジュールの価値体系が相互に影響しあって知識を再構成する作業である。

2-4-2. 自発性の導入

これまで考えてきたシステムは、基本的に受動的であったので、これに自発的な性質を与える。原始的なレベルでは、例えば空腹時には食べ物を求める行動をとるべきであるように、その内部的な欲求によってつき動かされる機能である。しかし、我々はより高度な自発的な行動を行なうができる、これは基本的な自発性が価値観と世界観の助けをかりて食べ物を買うためにはお金が必要だから自発的にお金を集めるなどのように原始的な自発性が拡張されたものである。よって、我々が行なう全て自発的な行動や思考は、根本的にはいくつかの基本的な欲求に還元できると思われる。

原始的な欲求による自発性を発現させるためには、例えば空腹時に認識連合モジュールにおいて食べ物などの関係するイメージを再現させるように、欲求を満足させることに関連するイメージを認識連合モジュール上に再現すれば良い。この作用を行なうモジュールを欲求モジュール(Desire module : DM)とする。欲求モジュールはかつてある欲求が満たさ

れたときの認識連合モジュールの出力を憶えておき、後に対応する欲求が起こった場合に、そのパターンを認識連合モジュール上に再現するように信号を出力する。

この機能を実現するためのモジュールの構造を説明するためにFig. 2-25に欲求モジュールを示すと同時にFig. 2-26にこれを含むニューロシステムの全体図を示した。欲求モジュールはシステムの内部条件に応じて例えば腹が減っていれば空腹信号をあるユニットに送るように設計された欲求発生器(Desire generator : DG)と、外界からの入力により満足時に出力を発生する快感発生器(Satisfaction generator : SG)（これは基本報酬発生器とほとんど同じであるが、機能的な意味で分けて表現した）および、快感発生器の出力によりある満足感に対応する認識連合モジュール内のパターンを学習するユニットを持つ欲求再現メモリー(Desire reappearing memory : DRM)により構成される。欲求再現メモリは複数個のユニットと学習促進器(LP)のペアから成り、あるユニットとペアを為す学習促進器が快感発生器の出力により刺激されたときに、認識連合モジュール上のパターンをそのユニットが覚える。すると後に欲求発生器からそのユニットに出力が送られて来たときにそのユニットが発火し、認識連合モジュールに以前満足感を感じたときと同じパターンが再現される。ここで、重要なことは欲求再現メモリにおける快感発生器からの満足信号と欲求発生器からの欲求信号は、満腹感と空腹感のペアのようにあらかじめ適切に結合されていることである。これは生物においては進化の過程において体得できる性質であろう。

このようにして本システムは自発性を得ることができ、生物のように外界からの刺激ではなく欲求を動機として自発的な行動や思考を行う事が可能となる。

ここでの、自発性の取扱は内的状態を特殊なものとして取り扱ったが、内的状態もシステムに対する入力と考えればこの様な新しい構成を付加しなくても原理的には実現できる、しかし内部的な情報の取扱については進化の過程でモデル化がある程度完成していると考えられるので、このように欲求モジュール(DM)の形で取り扱うのも悪くないだろう。

2-4-2. 好奇心の導入

人間は新しい物に対して興味を示す好奇心を持っている。もし新規性を示す信号を生成することができれば、この信号を基本報酬性価値の一つに加えればこの機能は実現できる。動作主体であるシステムにとっての新規性は、ある入力がこれまでの経験した入力と比較してどの程度異なっているかによる。しかもただ異なっているだけではなく、その入力がシステムの推論しえる予測に反する事が必要であろう。新規性信号を得るためには、認識連合モジュールの活動を観測すればよい。つまり、認識連合モジュールは経験を積んだ学習済みの入力パターンに対しては対応する安定点を持ちユニット出力の分布が明確にON

とOFFに分離すると期待されるが、未学習の入力パターンに対してはその様に振舞わないため、この両者の違いを区別できる。環境における適応という視点からは好奇心の効果は環境中の様々な入力の中から注意を与えるべき情報を、これまでの経験とできるだけ矛盾するものとした方が得られる情報量が大きくなるためだと思われる。よって、次々と新しい展開が起こる入力パターンは人間にとって面白いものとなる。だから若い人間が色々なものに興味を持つのは、まだ経験が少ないためにより多くの入力パターンによって新規性信号が発生するためだと思われる。また新規性は思考を行なう過程でシステム自身が生成できる報酬性価値観であるから、ある思考の結果得られた産物にも報酬性価値を与えることができる。そのため研究者の行なうような長期にわたる思考に対する動機付けを支える機能を果たせるだろう。

2-5. まとめ

本章では環境と相互作用しながら適応するシステムを基とたパターン情報に対するの知識処理を実現するシステムを提案した。パターン処理では対象と成る情報が膨大なので取り扱う範囲を絞り込むためになんらかの基準が必要である。そこで報酬性と嫌悪性の批判信号を基として重要性を含む価値体系を生成することを提案した。ここでは再帰的な構造により環境に適応した価値観を自動的に生成する適応的価値評価モジュールを開発した。さらに思考を環境より入手した情報をより取り扱い易い有効なものとする操作として位置付け、この操作は情報の相関を強調することと仮定した。そして、その作業をそれぞれ世界観と価値観をもつ二つのネットワークの共鳴によって行う方法を提案した。

多くの情報処理マシンではあらゆるデータに対する取扱が全く公平であるが、このモデルの認識連合モジュールで行なわれる思考に相当する情報処理は、むしろ偏見に満ちあふれた一人よがりのいい加減な推論が行なわれる。このような特性はある場合には欠点であるが、時に非常な長所となる。

本モデルの一つの意義はパターン型知能処理マシンに対する一つの全体像が与えられた点である。もちろんまだまだ検討の余地があるがこのモデルは一つの参考となるだろう。また、このモデルを含めてこの種のモデルが提案されるなら個々ののモジュールの必要とする機能が示唆されるのでネットワーク開発の目標も提示できる。

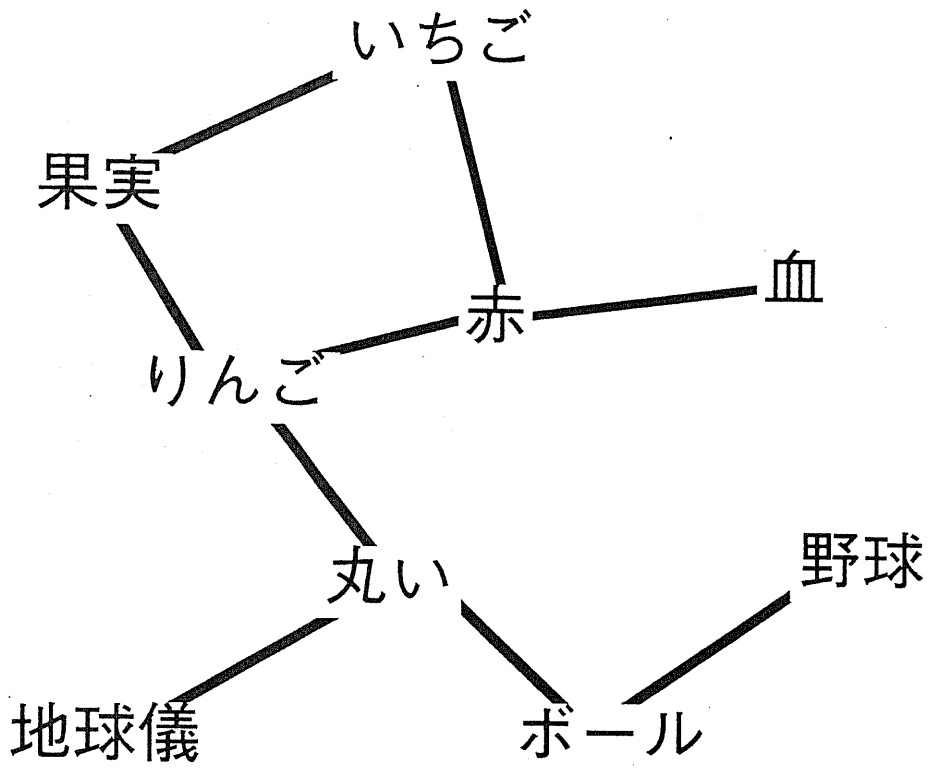


Fig. 2-1 相関に基づく知識

オブジェクト.関係.法則などは、区別できないし、基準も存在しない。境界のない相対的な関係のみが存在する

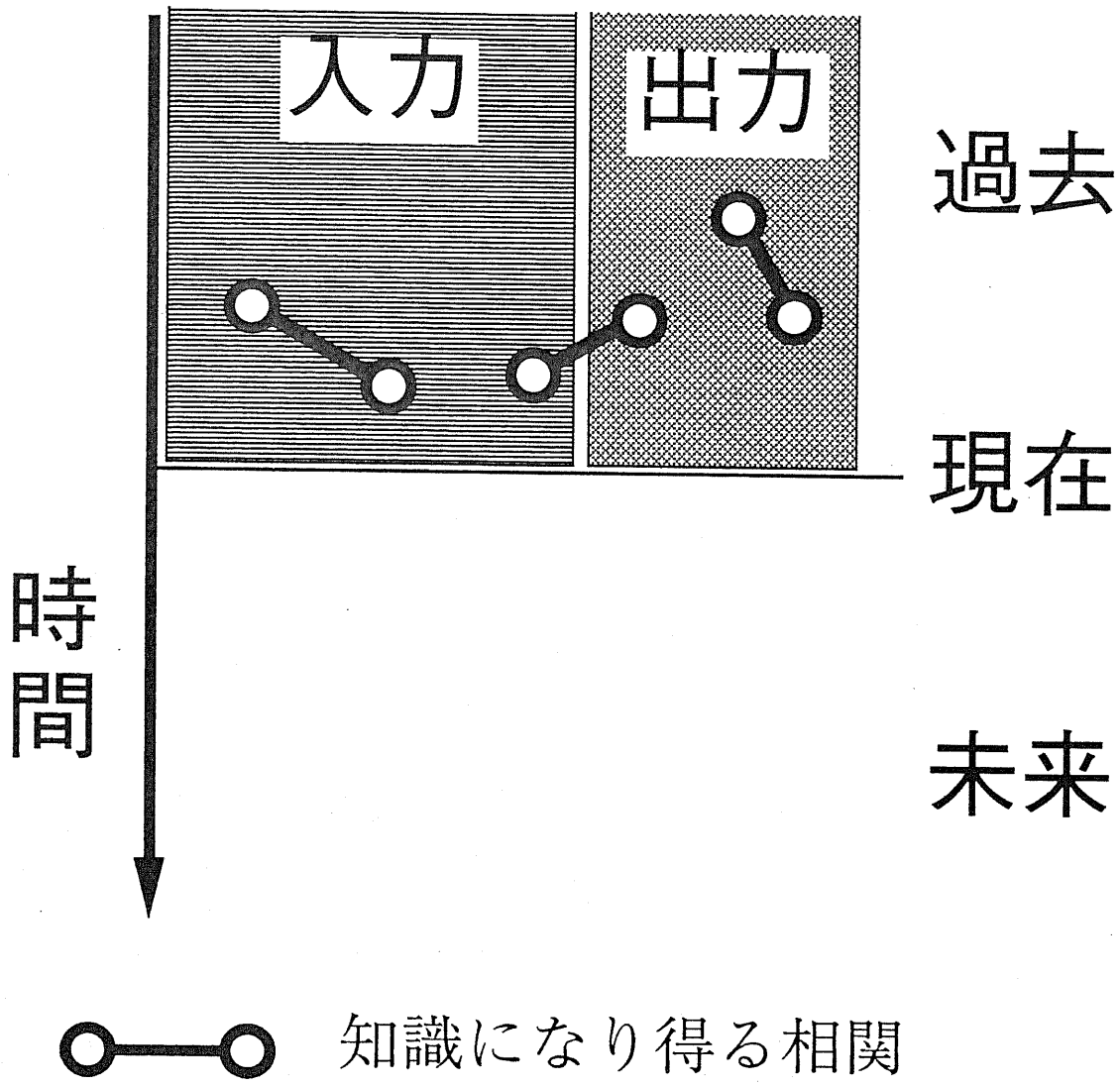


Fig. 2-2 動作主体が利用できる情報

入力.出力を含めた相関をが知識として利用し得る。

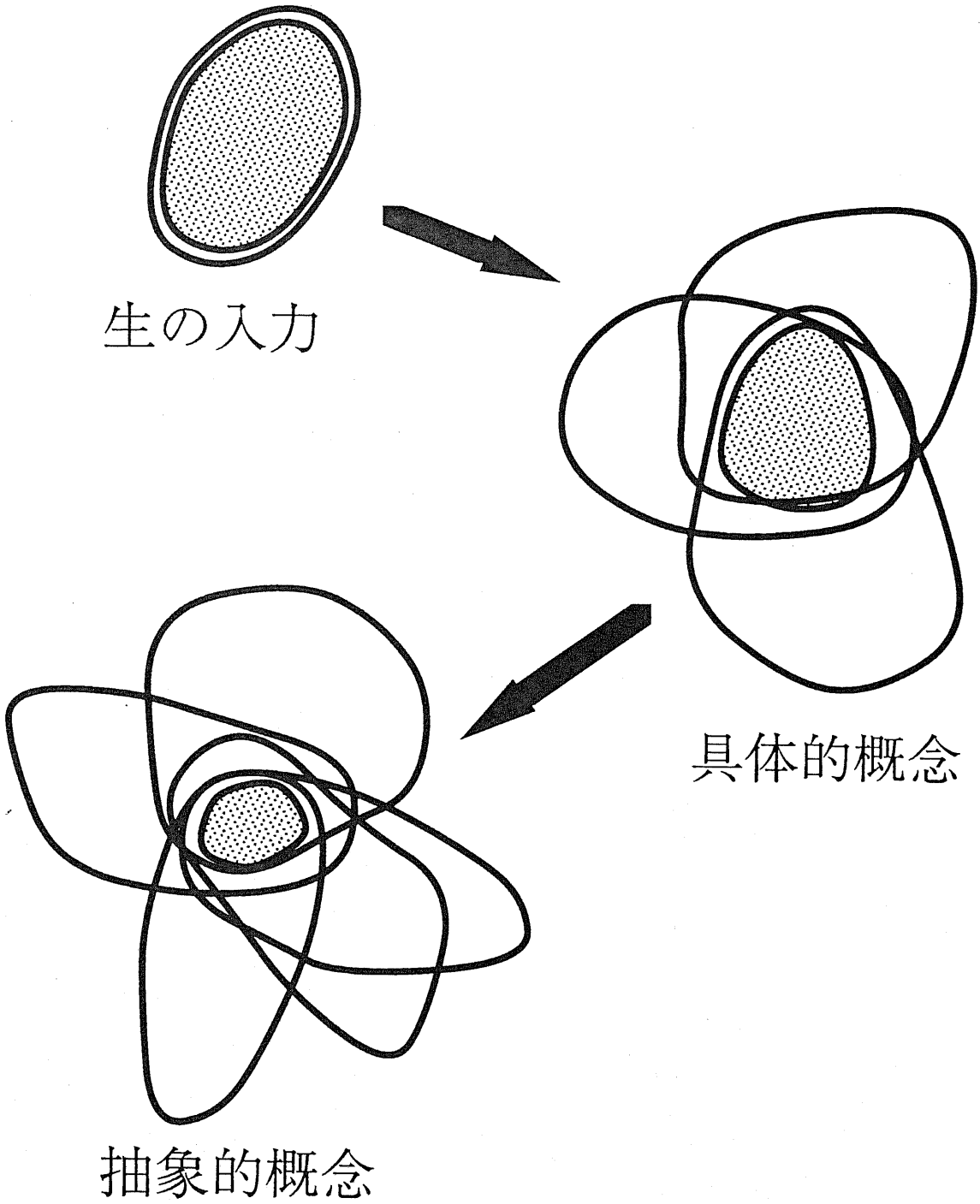


Fig. 2-3 抽象化のイメージ
一般性が高くなるにしたがって共通属性が減少し、予測能力の向上が期待される。

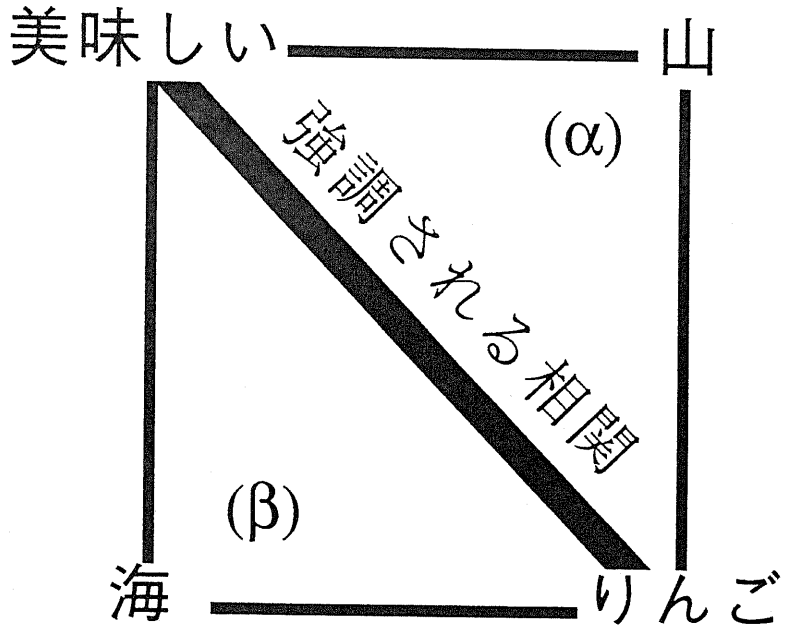


Fig. 2-4 思考による相関の強調

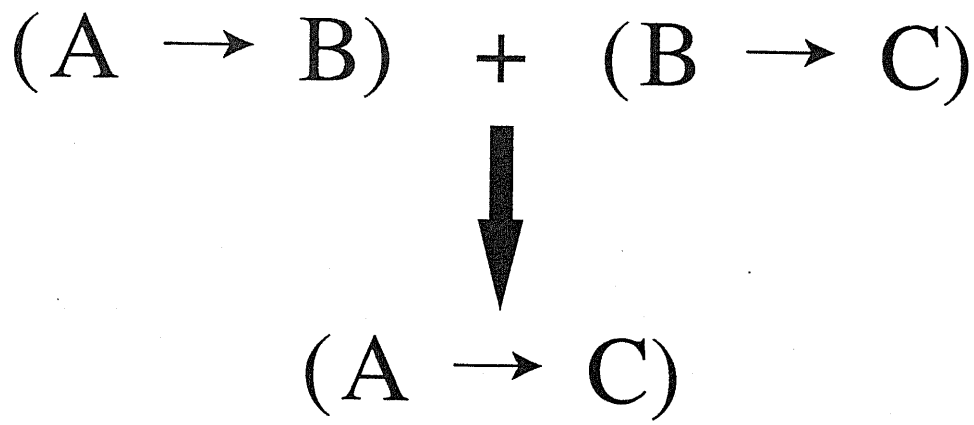


Fig. 2-5　思考による相関の結合

例えば、A：美味しい，B：リンゴ，C：お金.

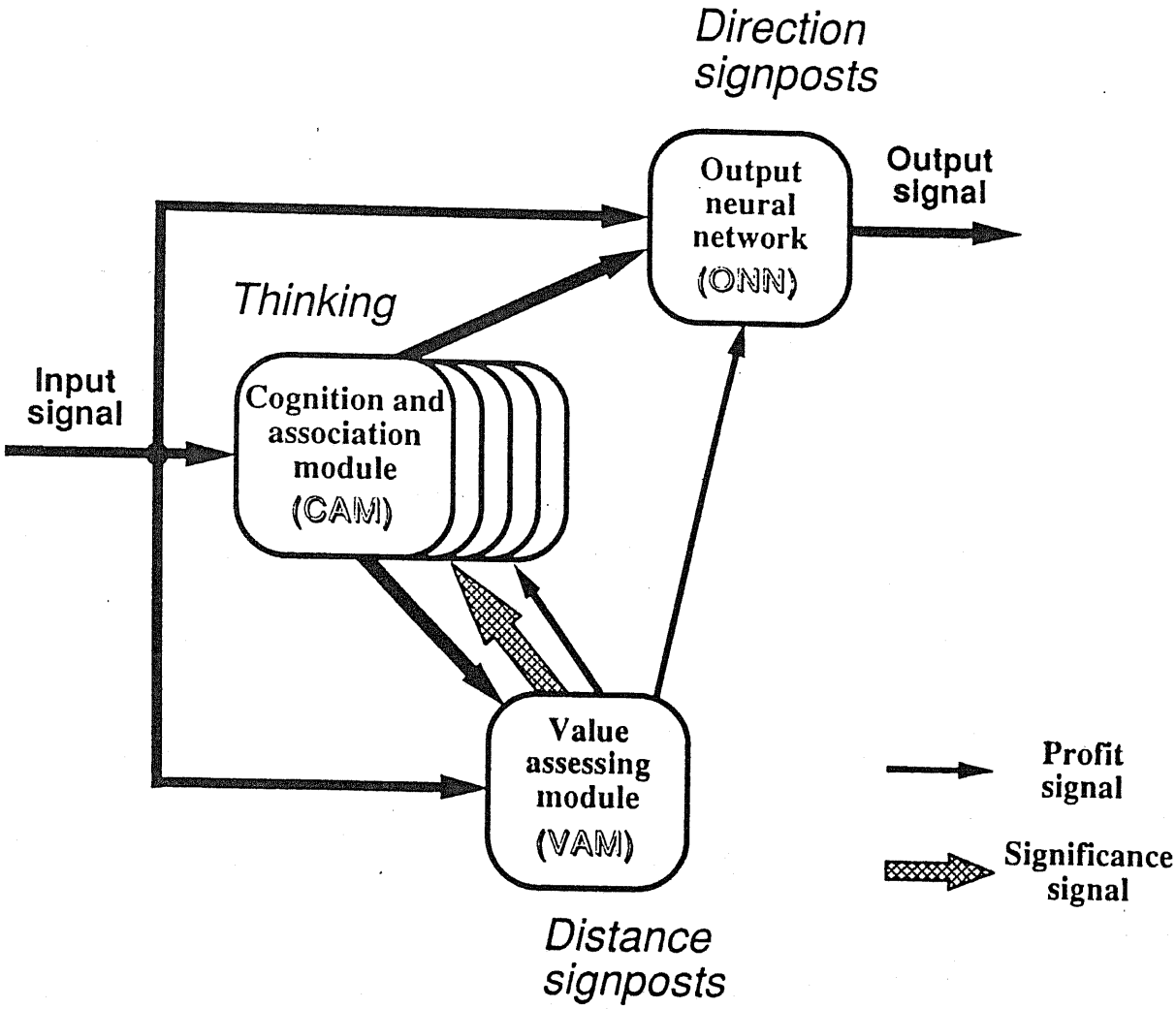


Fig. 2-6　動作主体の基本構成

Output neural network : 出力ニューラルネットワーク

Cognition association module : 認識連合モジュール

Value assessing module : 価値評価モジュール

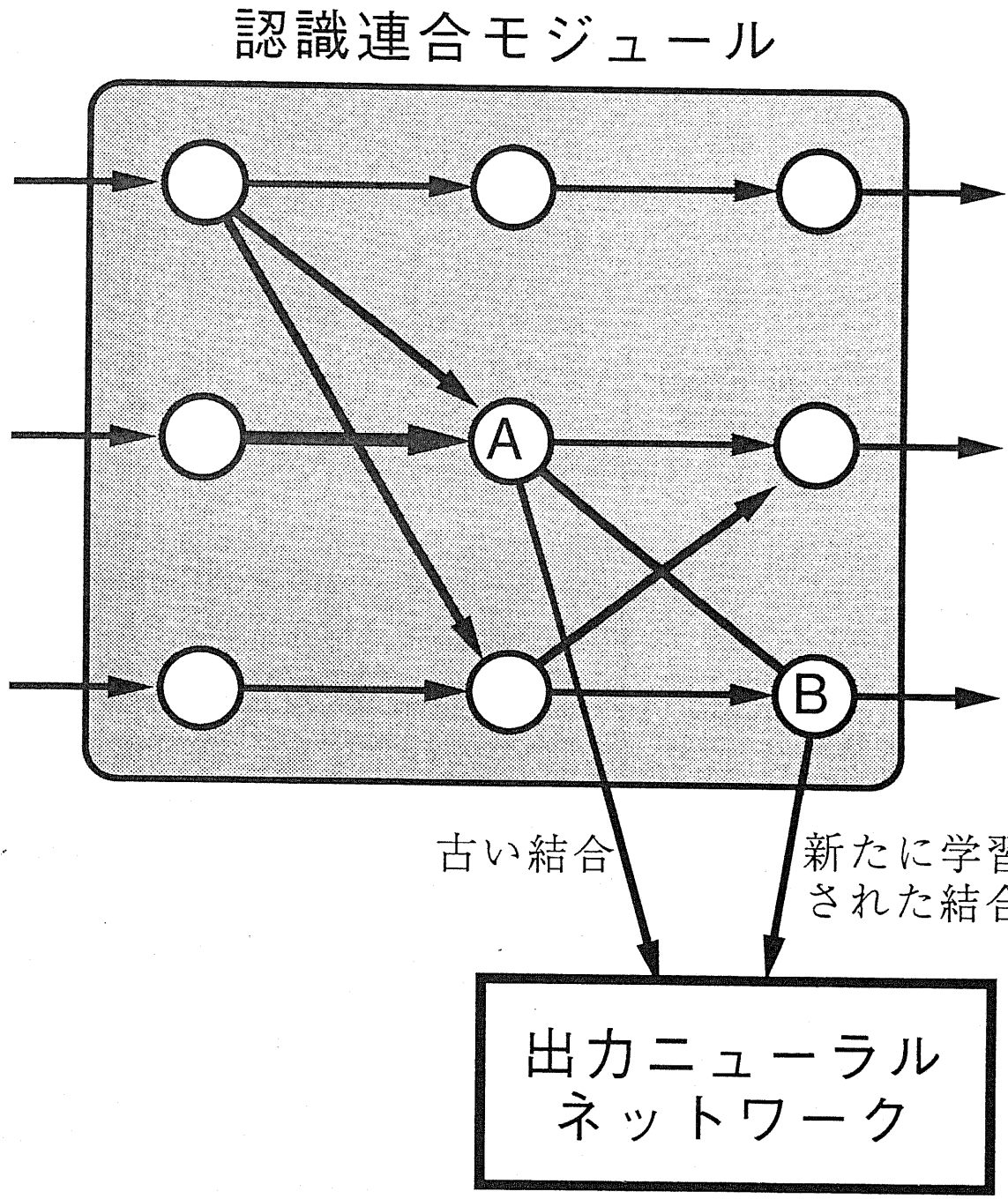


Fig. 2-7 認識連合モジュールの漸進的モデリング

Bの反応性を変更する際にもし ➡ の結合を変化させると、Aの反応性も変化するために好ましくない。（古い結合による動作が、不適切な反応性を持ってしまう）

そこで出力ネットワークと結合を持つユニットの入力結合はなるべく変化させない方がよい。

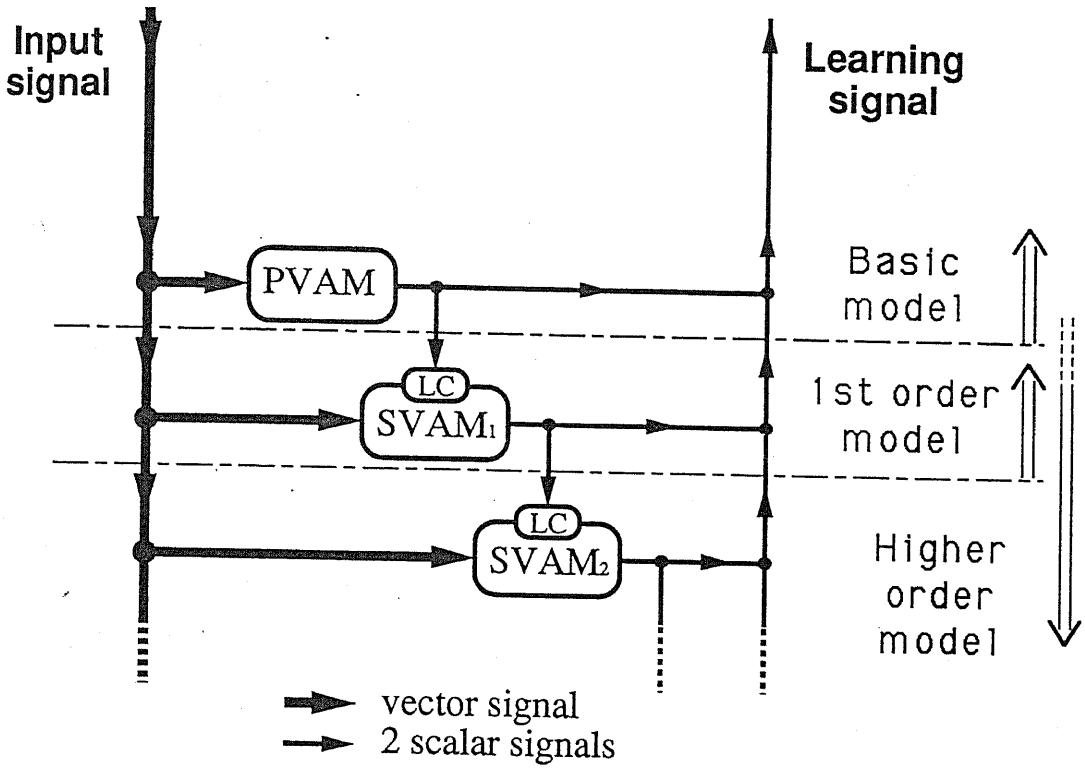


Fig. 2-8　単純な価値評価モジュール

価値評価のオーダーが二次価値評価モジュールのオーダーに対応している

PVAM : Primary value assessing module : 基本価値評価モジュール

SVAM : Secondaryvalue assessing module : 二次価値評価モジュール

LC: Learning controler : 学習制御器.

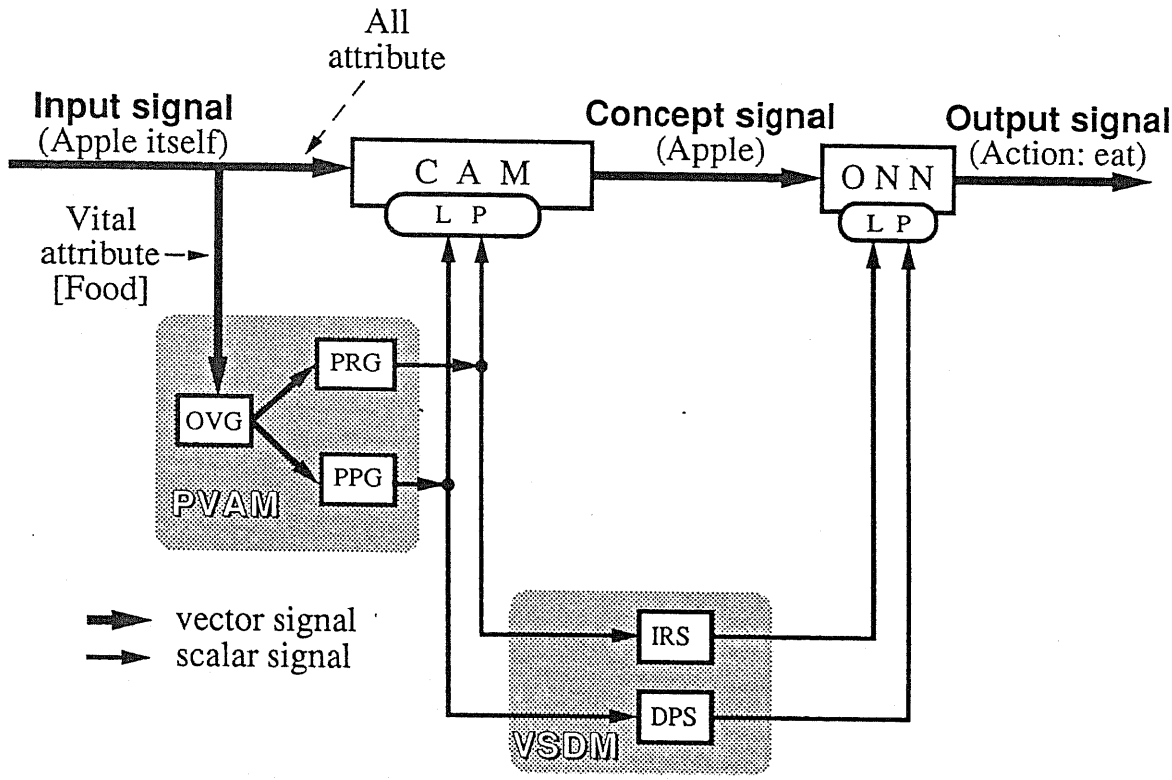


Fig. 2 - 9 The Primary Model of Neural System

How to answer basic needs like food.

ONN: Output neural network, CAM: Cognition and association module, LP: Learning prompter, PVAM: Primary value assesing module, OVG: Original value generator, PRG: Primary reward generator, PPG: Primary punishment generator, VSDM: Variation scen e detective module, IRS: Increasing reward sense detector, DPS: Decreasing punishment sen e detector,

ONN: 出力ニューラルネットワーク, CAM: 認識・連合モジュール, LP: 学習促進器, RLP: 反学習促進器,PVAM: 基本価値評価モジュール, OVG: 価値発生源, PRG: 基本報酬発生器, PPG: 基本嫌悪発生器, VSDM: 変化感知モジュール, IRS: 報酬増加感知器, DPS: 嫌悪減少感知器.

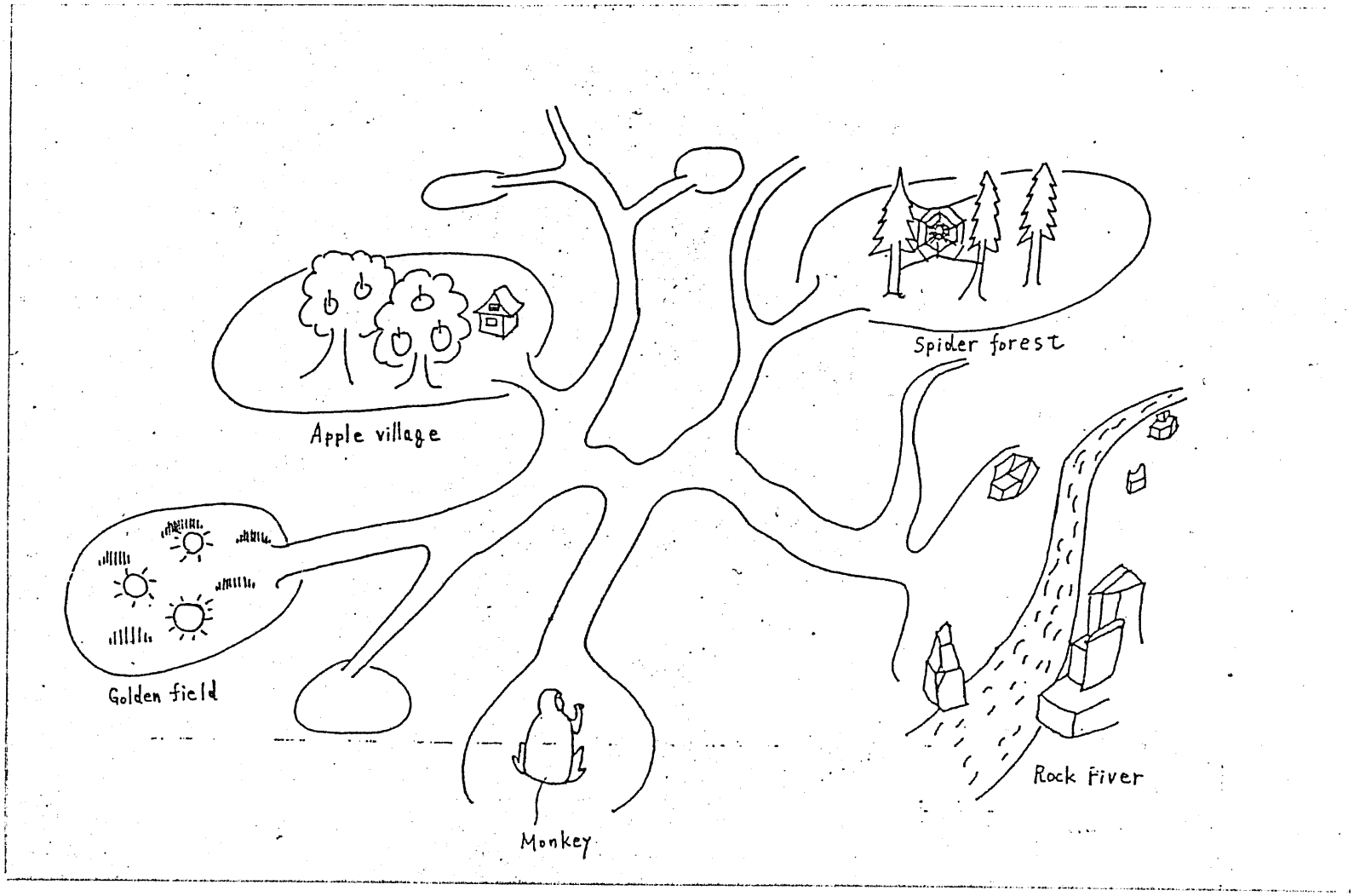
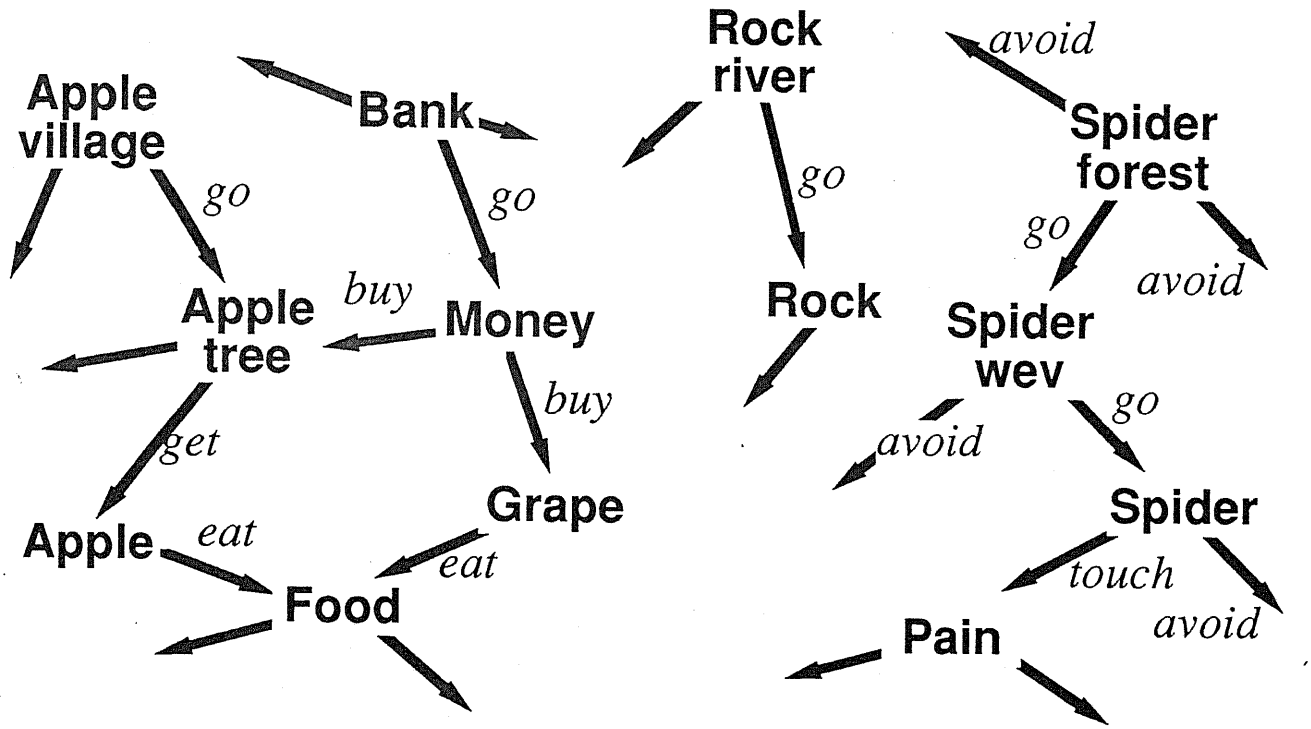
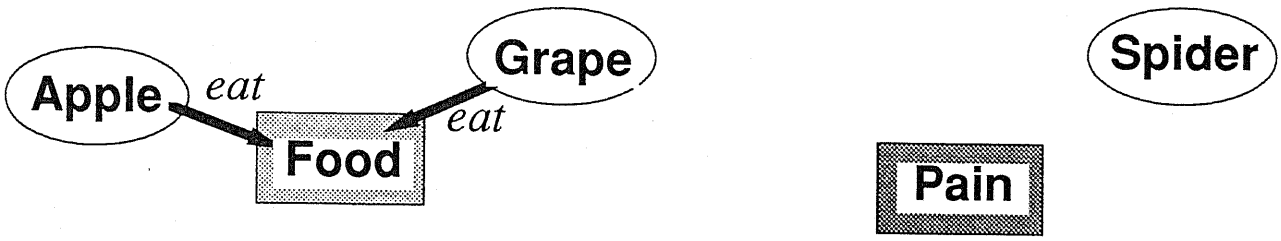


Fig. 2-10　シミュレーションに利用した環境のイメージ



Bold words mean input states.
Italic words mean actions.

Fig. 2-11 シミュレーションに使用した環境の関係図



object (categorized output)[CAM],
primary value [PVM],
reward value,
punishment value.

Bold words mean input states.
Italic words mean actions.

Fig. 2-12　基本モデルの学習能力

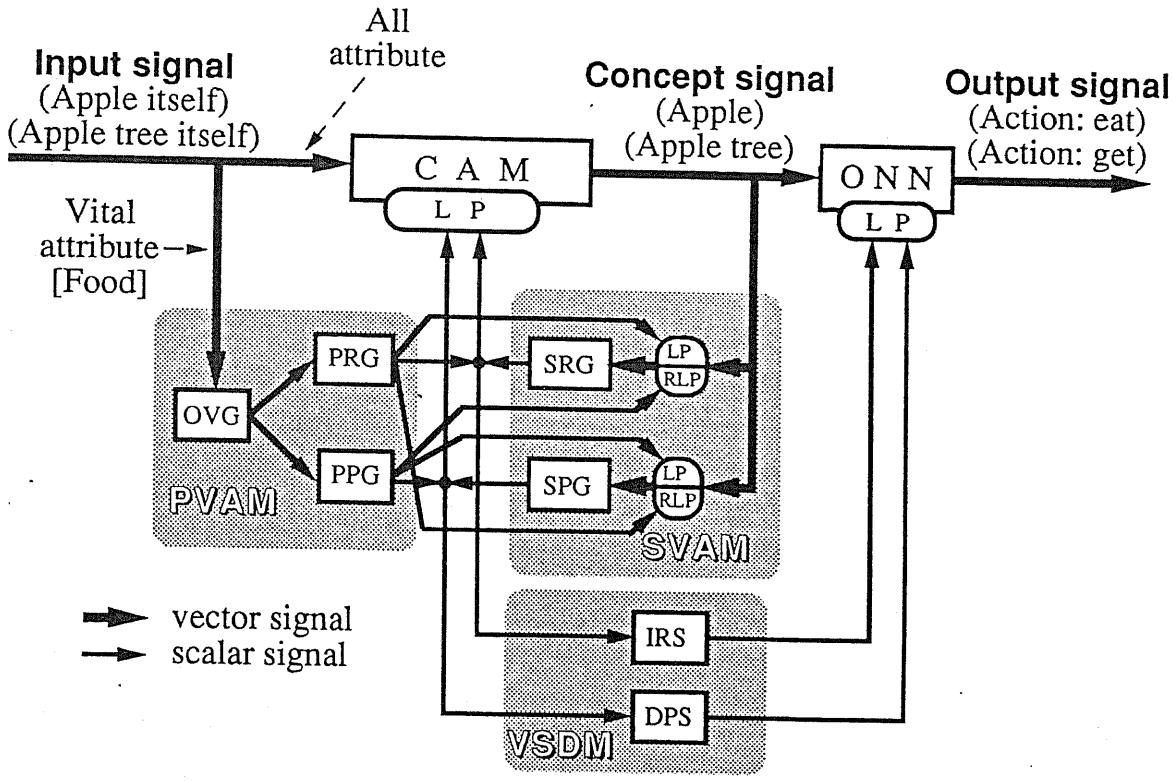
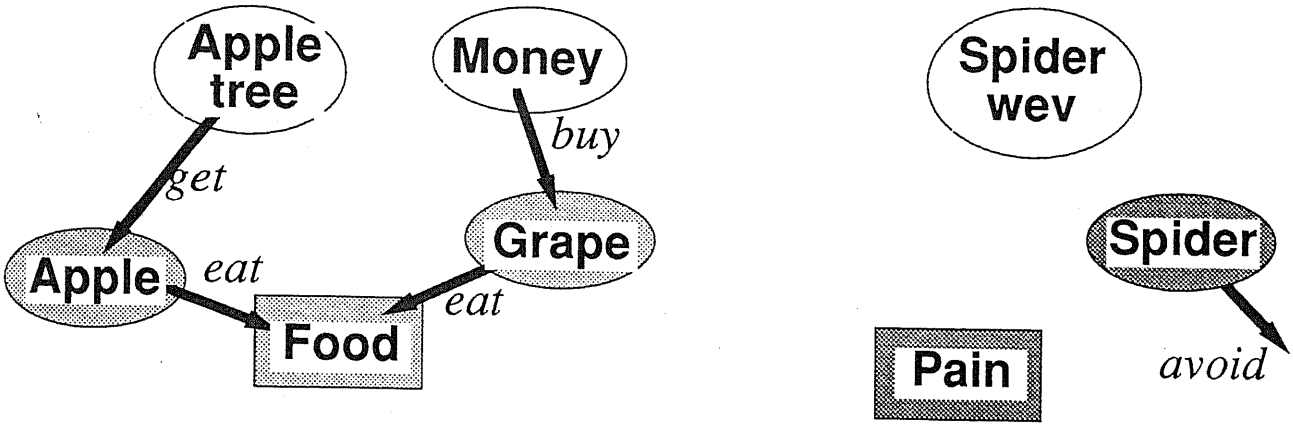


Fig.2-13 First Modelization of The General Answer of Neural System

How to answer primary but more general need.

ONN: Output neural network, CAM: Cognition and association module, LP: Learning prompter, PVAM: Primary value assesing module, OVG: Original value generator, PRG: Primary reward generator, PPG: Primary punishment generator, VSDM: Variation scence detective module, IRS: Increasing reward sense detector, DPS: Decreasing punishment sence detector, SVAM: Secondary value assesing module, SRG: Secondary reward generator, SPG: Secondary punishment generator, RLP: Reverse learning prompter,

ONN: 出力ニューラルネットワーク, CAM: 認識・連合モジュール, LP: 学習促進器, RLP: 反学習促進器,PVAM: 基本価値評価モジュール, OVG: 価値発生源, PRG: 基本報酬発生器, PPG: 基本嫌悪発生器, VSDM: 変化感知モジュール, IRS: 報酬増加感知器, DPS: 嫌悪減少感知器, SVAM: 二次価値評価モジュール, SRG: 二次報酬発生器, SPG: 二次嫌悪発生器.



object (categorized output)[CAM],
primary value [PVM],
reward value,
punishment value.

Bold words mean input states.
Italic words mean actions.

Fig. 2-14 一次モデルの学習能力

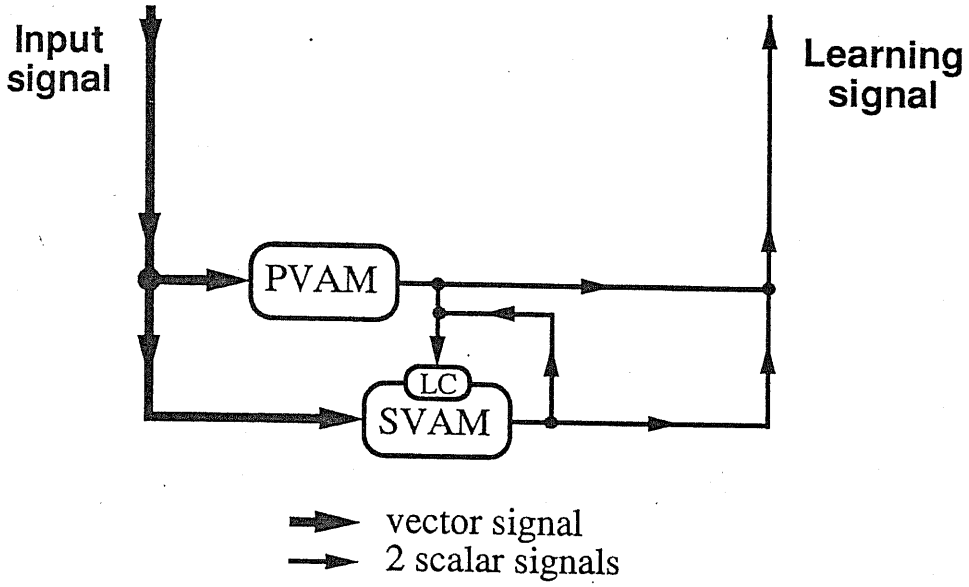


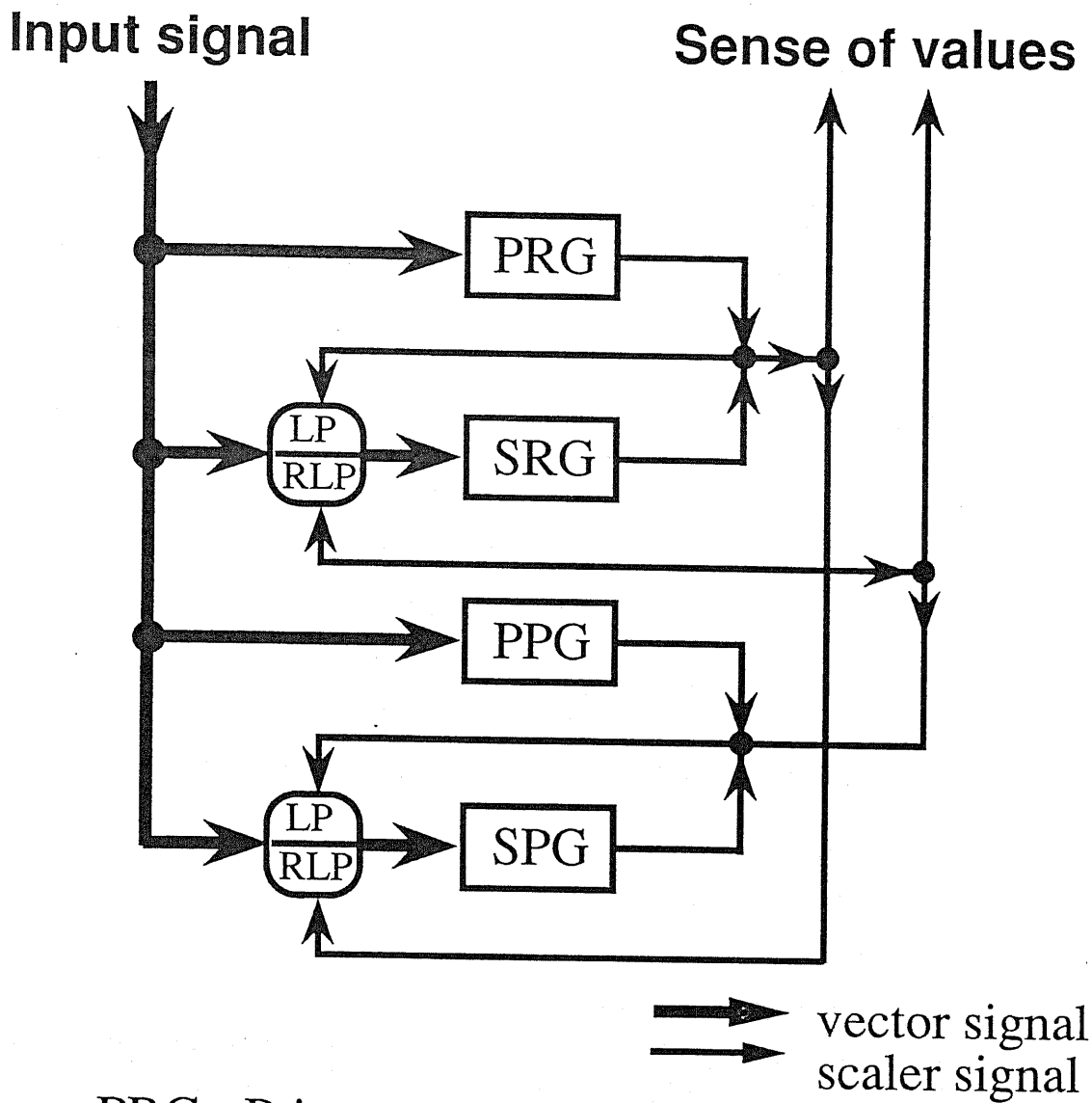
Fig. 2-15　再帰的な価値評価モジュール

全てのオーダーの二次価値が同一つの二次価値評価モジュール内に記憶される。

PVAM : Primary value assessing module : 基本価値評価モジュール

SVAM : Secondaryvalue assessing module : 二次価値評価モジュール

LC: Learning controler : 学習制御器.



PRG: Primary reward generator,
PPG: Primary punishment generator,
SRG: Secondary reward generator,
SPG: Secondary punishment generator,
LP: Learning promoter,
RLP: Reverse learning promoter.

Fig. 2-16　再帰的な価値評価モジュールの詳細

報酬性と嫌悪性を分けて表現した。

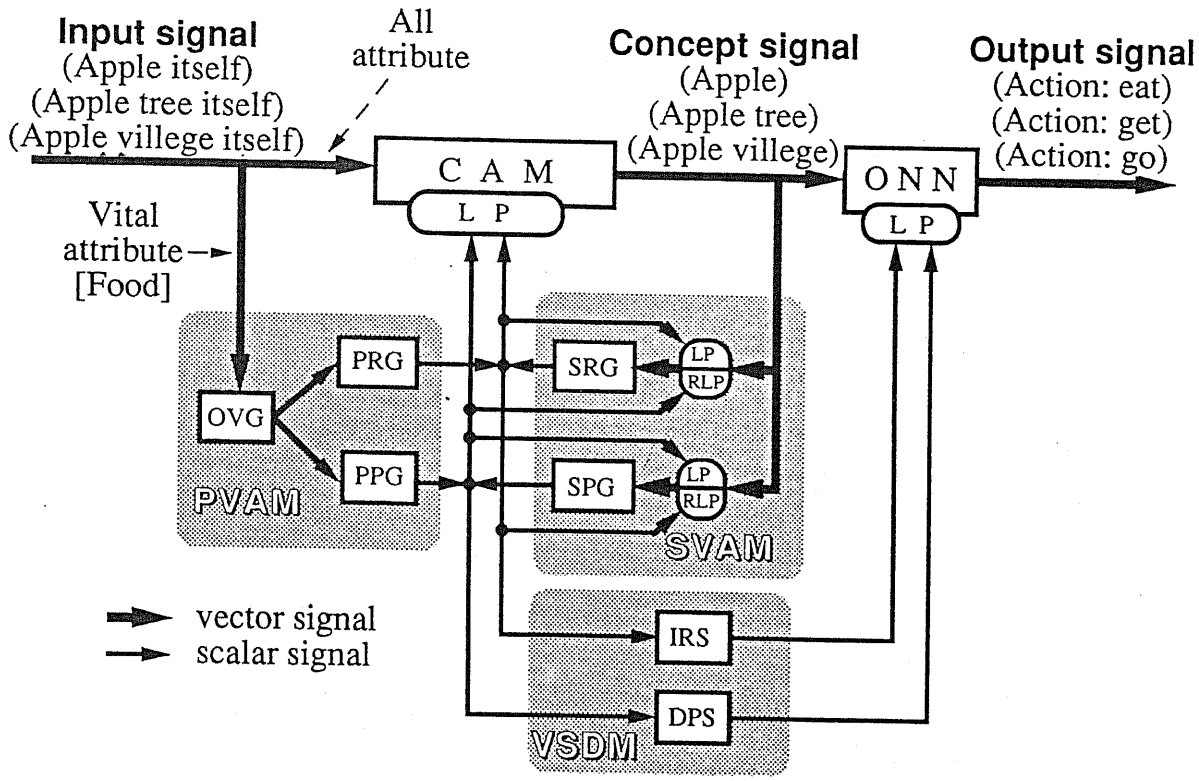
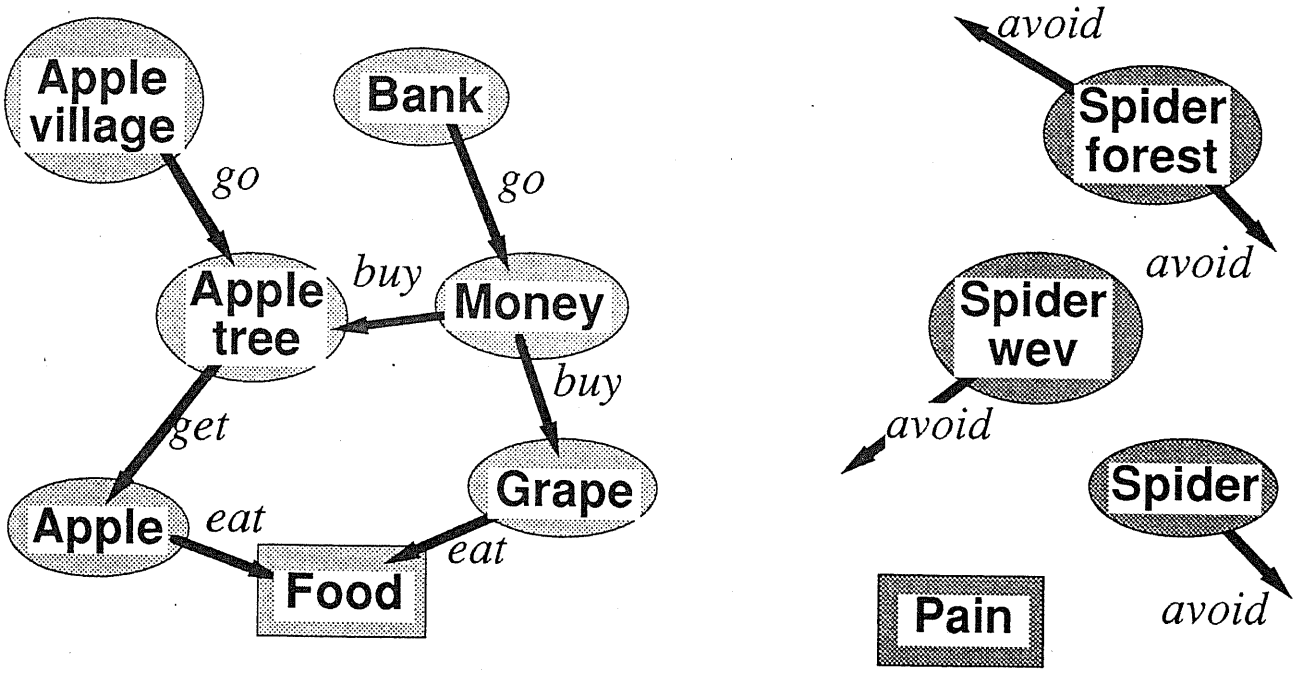


Fig. 2-17 The Model of Recursive Neural System

How to answer general need.

ONN: Output neural network, CAM: Cognition and association module, LP: Learning prompter, PVAM: Primary value assesing module, OVG: Original value generator, PRG: Primary reward generator, PPG: Primary punishment generator, VSDM: Variation scence detective module, IRS: Increasing reward sense detector, DPS: Decreasing punishment sence detector, SVAM: Secondary value assesing module, SRG: Secondary reward generator, SPG: Secondary punishment generator, RLP: Reverse learning prompter,

ONN: 出力ニューラルネットワーク, CAM: 認識・連合モジュール, LP: 学習促進器, RLP: 反学習促進器,PVAM: 基本価値評価モジュール, OVG: 価値発生源, PRG: 基本報酬発生器, PPG: 基本嫌悪発生器, VSDM: 変化感知モジュール, IRS: 報酬増加感知器, DPS: 嫌悪減少感知器, SVAM: 二次価値評価モジュール, SRG: 二次報酬発生器, SPG: 二次嫌悪発生器.



object (categorized output)[CAM],
primary value [PVM],
reward value,
punishment value.

Bold words mean input states.
Italic words mean actions.

Fig. 2-18 再帰モデルの学習能力

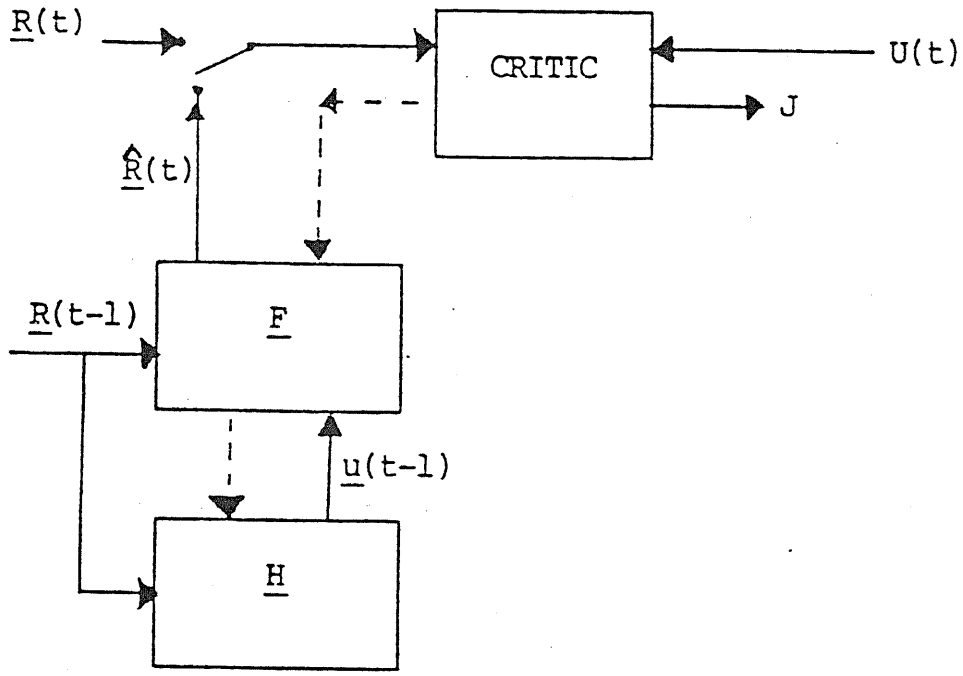


Fig. 2-19　Werbos のモデル

Architecture for Backpropagated Critic

(J-base versin; λ -based is similar) (Werbos, 1989)

H：動作ネットワーク,　　F：エミュレータ・ネットワーク,

R：状態,　　u：コントロールシグナル,

U：現在の状態に対する評価,　J：二次的（戦略的な）評価.

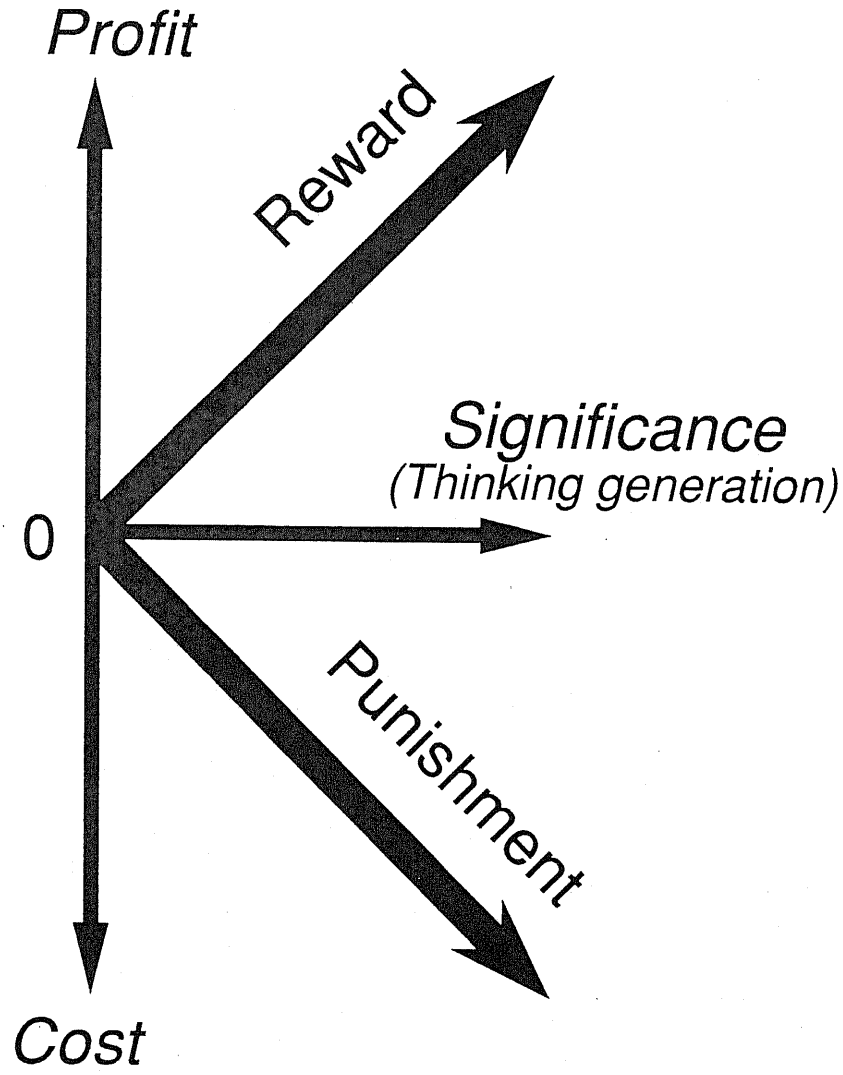


Fig. 2-20　価値観の次元

本来の次元は報酬性(Reward)と嫌悪性(Punishment)であるが、使用される次元は重要性(Significance)と評価(Profit/Cost)の次元である。

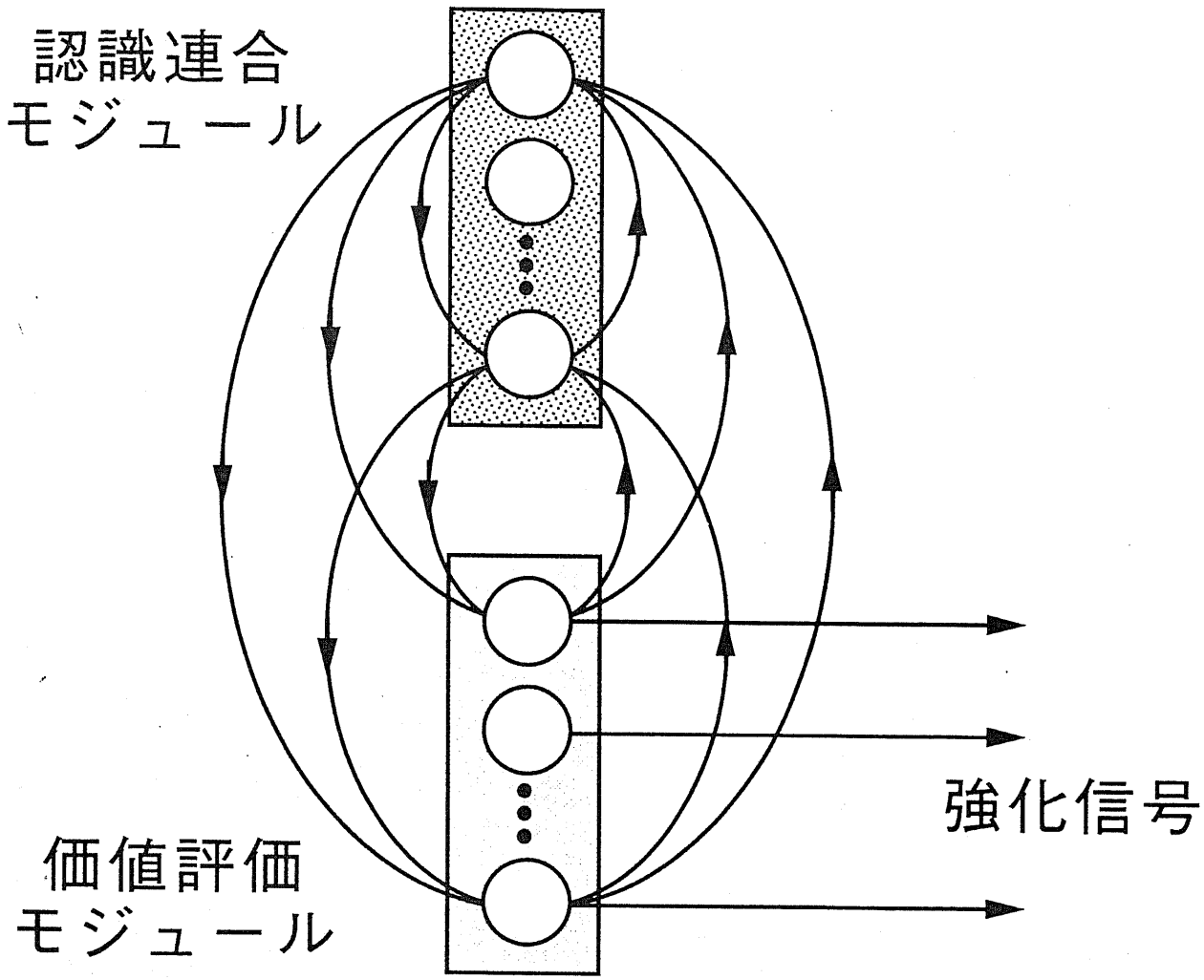


Fig. 2-2l 認識連合モジュールと価値評価モジュールの相互作用

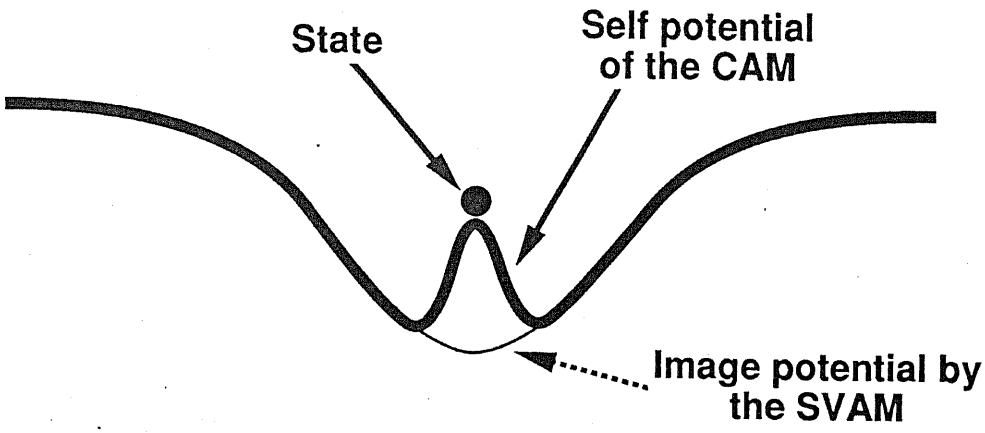


Fig. 2-22　認識連合モジュールの状態に働くポテンシャル

認識連合モジュールの状態空間における状態の位置をドットで表している。状態に対して連想のための速度の速い斥力型のポテンシャルと、意識に対応する価値評価モジュールの出力による引力型のポテンシャルが働いている。

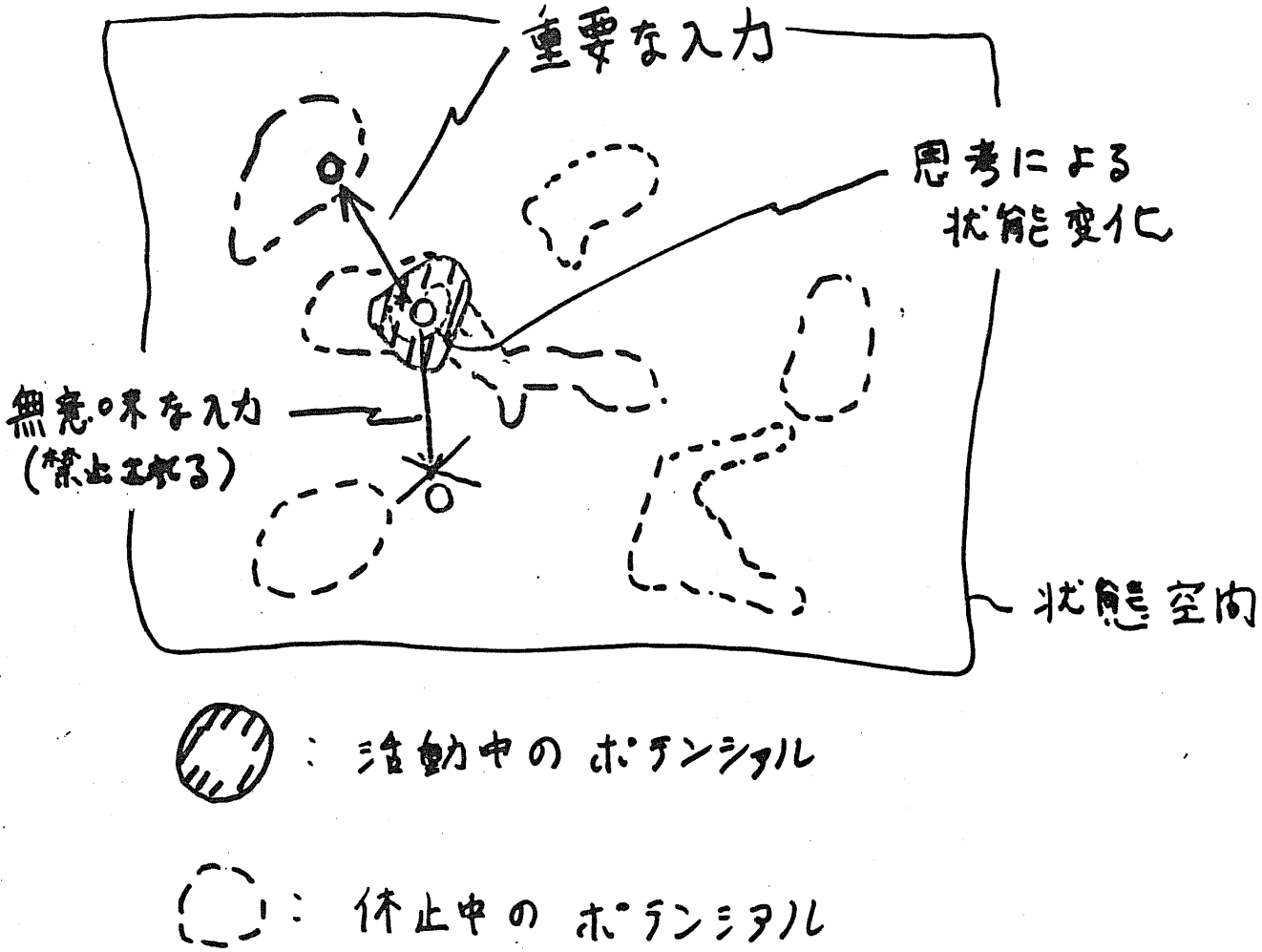


Fig. 2-23思考と認識連合モジュールの状態空間

状態空間中のポテンシアルは、価値評価モジュールによってマッピングされる。

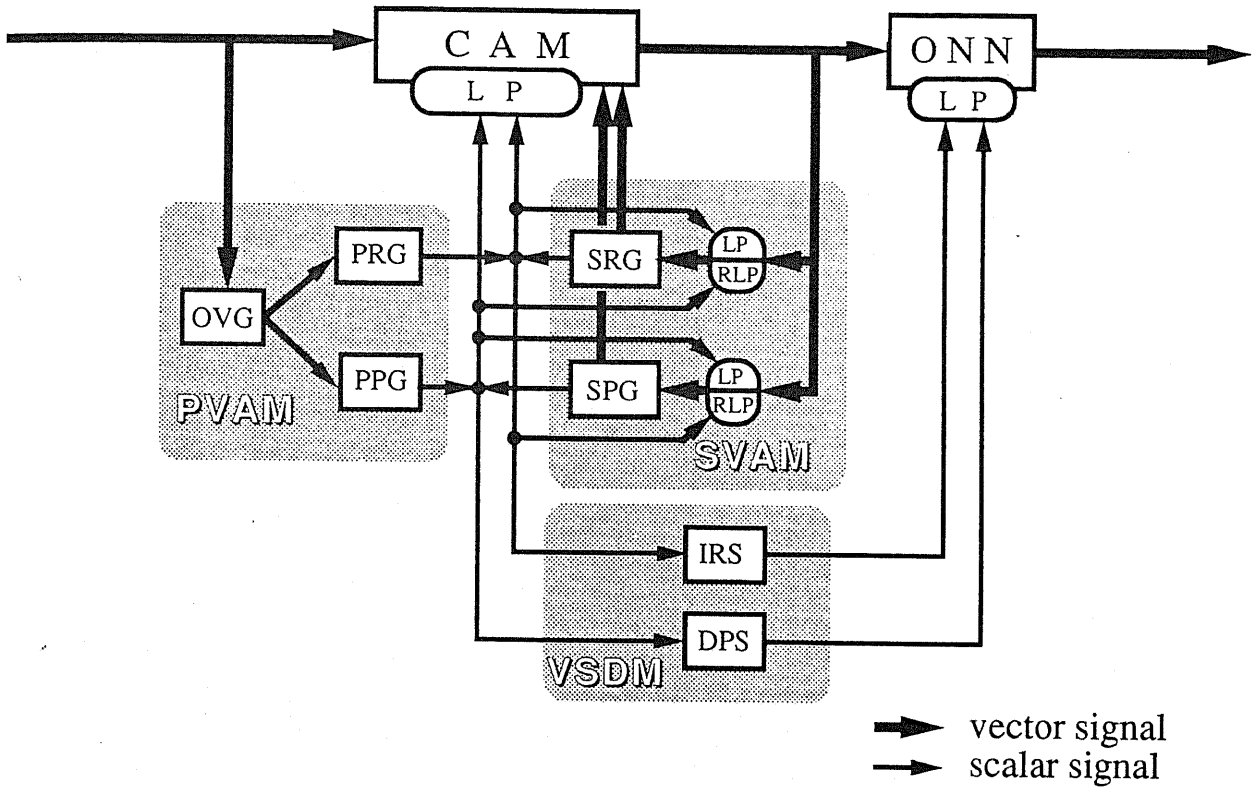


Fig.2-24　思考するニューラルシステム

ONN: Output neural network, CAM: Cognition and association module, LP: Learning promoter, PVAM: Primary value assesing module, OVG: Original value generator, PRG: Primary reward generator, PPG: Primary punishment generator, VSDM: Variation scence detective module, IRS: Increasing reward sense detector, DPS: Decreasing punishment sence detector, DM: Desire module,

ONN: 出力ニューラルネットワーク, CAM: 認識・連合モジュール, LP: 学習促進器, RLP: 反学習促進器,PVAM: 基本価値評価モジュール, OVG: 価値発生源, PRG: 基本報酬発生器, PPG: 基本嫌悪発生器, VSDM: 変化感知モジュール, IRS: 報酬増加感知器, DPS: 嫌悪減少感知器, SVAM: 二次価値評価モジュール, SRG: 二次報酬発生器, SPG: 二次嫌悪発生器, DM: 欲求モジュール, SG: 快感発生器, DG: 欲求発生器,DRM: 欲求再現メモリ.

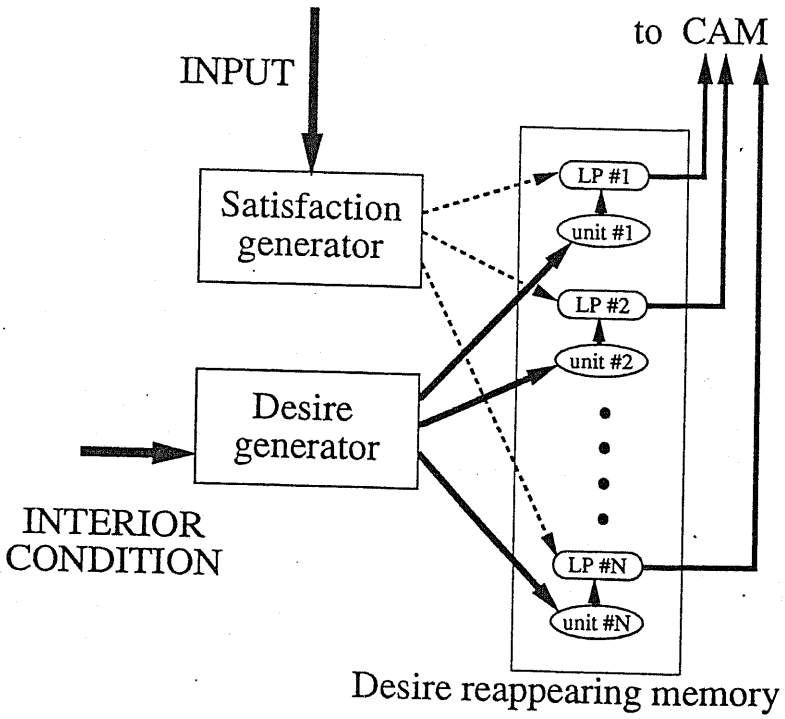


Fig. 2-25　欲求モジュール(Desire moduel : DM)

Satisfaction generator (SG) :　快感発生器,

Desire generator (DG) :　欲求発生器,

Desire reappearing memory (DRM) :　欲求再現メモリ.

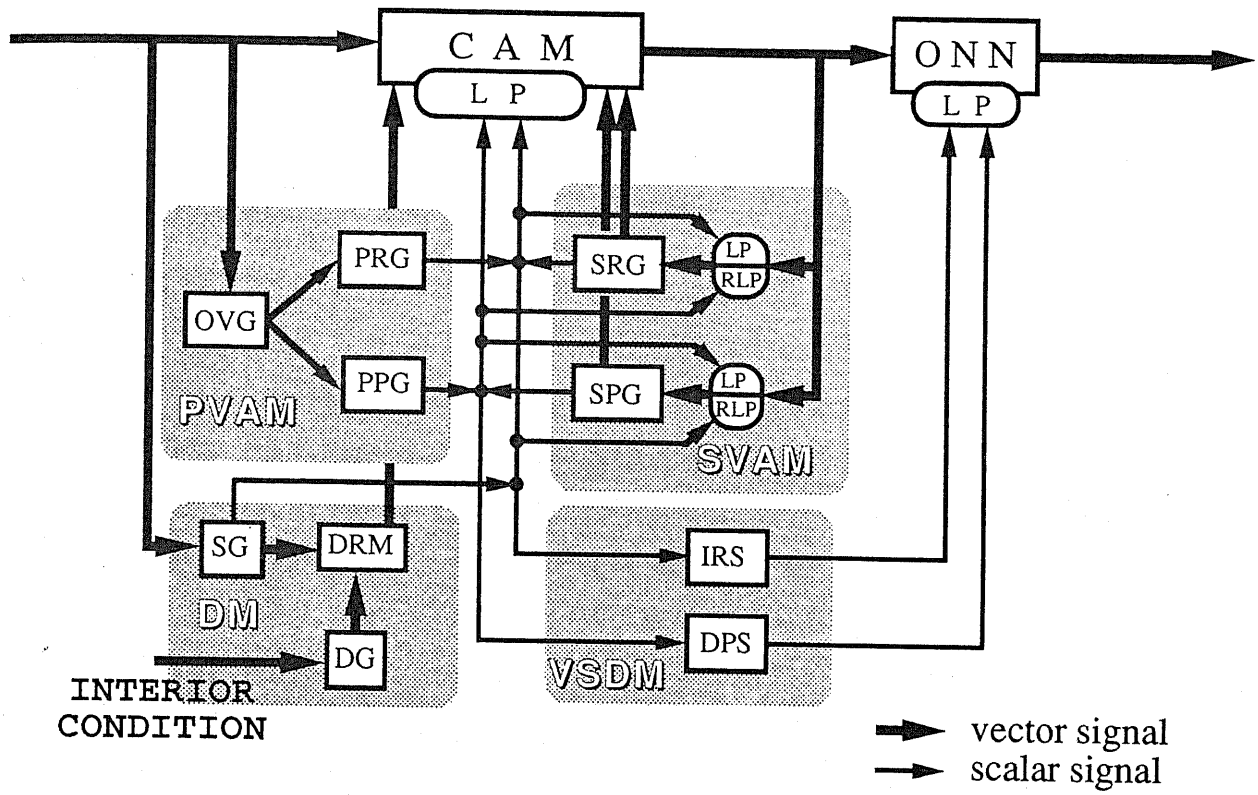


Fig.2-26　自発性を備えたニューラルシステム

ONN: Output neural network, CAM: Cognition and association module, LP: Learning promoter, PVAM: Primary value assesing module, OVG: Original value generator, PRG: Primary reward generator, PPG: Primary punishment generator, VSDM: Variation scence detective module, IRS: Increasing reward sense detector, DPS: Decreasing punishment sence detector, DM: Desire module, DG: Desire generator, SG : Satisfaction generator, DRM : Desire reappearing memory.

第3章

ネットワークの状態制御

本章では相互結合型ネットワークにグローバルな状態を制御する機構を付加することを提案する。ここで制御すべき状態とは活動状態における出力の分布である。出力の分布状態を表す出力レベル密度(DOOL)関数を定義し、この形を制御する。

ニューラルネットワークにおいて標準的なしきい値素子等の一種類のユニットにより構成される相互結合型ネットワーク上の結合は情報保持と状態制御の二つの機能を果たす必要がある。そこで、状態制御を他の機構に任せればネットワーク上の結合は情報を貯える機能（学習）に専門化でき、その能力の向上が期待される。

状態制御は高速な反応を要求する代わりにユニット出力の平均値等の単純な情報で充分で、ユニット間の結合毎の複雑な制御を行う必要がない、よって簡単な他の構造を用いるのが賢明である。一方、情報保持はユニット間の複雑な情報を取り扱う代わりに速い反応は必要としない、よって結合の変化に任せるのが適切である。

この方法を用いると学習を行っていない状態のネットワークでも尤もらしい出力を行う。多くの中間ユニットを含むネットワークでは、冗長なユニットは尤もらしい出力を行う事によりそれなりの結合が形成され、入力に対する新たな区別が要求された際に柔軟な対応が期待できる。

出力レベル密度関数を一定の形に制御するとネットワークが表現できる出力パターンは制限される。しかしネットワーク内のユニットの数が増えた場合その処理能力は表現できる出力パターンの複雑さに比べて小さくなるので、この制限はネットワークのパフォーマンスを低下させず、適当な状態制御は情報処理能力を向上させることができると思われる。またここでは隠れユニット多数含むネットワークについて考えているので、ネットワーク内の部分集合ユニット郡である入力ユニット郡や出力ユニット郡は全ての状態が実現できる。

3-1. 望ましい出力状態

活動時の出力レベル密度関数の形に対する要求はシステムの使用目的により異なる。しかしどの様な使用目的であってもその関数の形は多数の状態を含むものでなくてはならない(宮下,1989)。

ここではさらに以下の二つの条件を課した。一つ目はユニットの出力が大まかに二値に分離していてＯＮとＯＦＦの出力を区別できること、二つ目は出力レベル密度関数がある程度広い範囲に分布していて大まかには類似の出力状態でも細かくは異なった出力状態が実現できることである　。すると必要な出力レベル密度関数はFig.3-1の様なひょうたん型となる。大まかにＯＮとＯＦＦの出力が区別できると、最大値を出力するユニットに関する結合だけ強化するなどの不自然な学習規則を設定せずとも学習時に強化すべき結合が区別できる。

3-2. ネットワークモデル

3-2-1. モデルの特徴

情報処理を行うネットワークは興奮性ユニットにより構成されるネットワークとした。これは抑制性の結合を含むネットワークの活動度変換関数は減少部分を含む複雑な特徴を持つのに対し、興奮性の結合のみの場合には活動度変換関数はFig.3-2に示すようなＳ字形の単調増加関数であり、このため状態を制御する付加機構は抑制性のみで足りるからである。ここで、我々は興奮性ネットワークの活動度が適当な値（活動値）になるとそれ以上活動度が増大しないよう抑制をかける機構を付加する事にした。これにより安定な活動値は前記活動値とほとんんど０の二つの状態である(伊藤,1989)。

一方、出力レベル密度関数の細かい制御は3-4.の中で説明するユニット毎の自己帰還入力により可能となる。

3-2-2. モデルの構成

次に、このモデルの具体的な構造についてFig.3-3を用いて説明する。情報処理を行う興奮層はその内部のユニットが互いに興奮性結合を持つとともに抑制層に対する出力として興奮性結合を持つ。この興奮層から抑制層への結合は平均的で特定のユニット間に強い結合は存在しない。他方抑制層は興奮層に対する出力として抑制性結合を持ち、抑制層内に相互結合は存在しない。そして、抑制層の動作速度と情報処理層である興奮層の動作速度はそれぞれ変化させることができる。それぞれの層内のユニットの出力範囲を０から１と規格化したので活動度（層内のユニット出力の平均値）の範囲も０から１である。

以下、興奮層と抑制層の活動度の関係を記述する方程式は、

$$\frac{dx}{dt} = \frac{1}{T_1}[F_1(x - \alpha y) - x]$$

$$\frac{dy}{dt} = \frac{1}{T_2}[F_2(x) - y] \qquad (3\text{-}1),(3\text{-}2)$$

x, y : 興奮層と抑制層の活動度, α : 定数,

T1, T2 : 興奮層と抑制層の時定数, t : 時間,

F1(), F2() : 興奮層と抑制層の活動度変換関数

となる(甘利,1978)。

3-3. 活動度を用いた研究

3-3-1. 活動度を用いた研究における制御目標

この段階でのシステムの実現すべき目標は 1)活動安定点が存在する事、2)静止安定点が存在する事、3)活動安定点での抑制層の出力が興奮層の出力に比べて小さい事、4)活動安定点に収束する入力の範囲（情報レンジ）が適当である事、の４つである。3)の要請は、抑制層の活動度が大きくなり興奮層の本来の信号成分に対するノイズ成分が大きくなるのを避けるためであり、4)の要請に関しては後ので詳しく述べる。

3-3-2. 線形近似による平衡点の解析

まず、興奮層と抑制層の相互作用により形成される平衡点の近傍での動的な振る舞いを考察するために、極めて微小な領域では関数 $F_1(x)$, $F_2(x)$ が線形近似できるものと考える。そして、平衡点における関数 $F_1(x)$, $F_2(x)$ の傾きを k_1, k_2 とした。(3-1),(3-2)式を線形化した後、平衡点を原点とし、行列の形で書き下すと、

$$\frac{d}{dt}\begin{pmatrix} x \\ y \end{pmatrix} = \frac{1}{2\pi}\begin{pmatrix} (k_1 - 1)\omega_1 & -\alpha k_1 \omega_1 \\ k_2\omega_2 & -\omega_2 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} \tag{3-3}$$

$$k_i = dF_i(\xi)/d\xi, \quad \omega_i T_i = 2\pi \ (i = 1, 2) \tag{3-4}$$

となる。

この固有値問題を x, y について解くと

$$x = \exp\left(\frac{P}{4\pi}\right)\left[C_+\exp\left(\frac{Q^2 - R^2}{4\pi}t\right) + C_-\exp\left(\frac{-Q^2 - R^2}{4\pi}t\right)\right]$$
$$y = \exp\left(\frac{P}{4\pi}\right)\left[D_+\exp\left(\frac{Q^2 - R^2}{4\pi}t\right) + D_-\exp\left(\frac{-Q^2 - R^2}{4\pi}t\right)\right] \tag{3-6),(3-7}$$

ここで、C+, C− は積分定数

$$D_+ = C_+\frac{Q - Q^2 - R^2}{S}, \qquad D_- = C_-\frac{Q - Q^2 - R^2}{S}$$

$$P = (k_1 - 1)\omega_1 - \omega_2, \qquad R^2 = 4\alpha k_1 k_2 \omega_1 \omega_2 \ (>0)$$
$$Q = (k_1 - 1)\omega_1 - \omega_2 \ (>0), \qquad S = 2\alpha k_1 \omega_1 \ (>0)$$

である。

この解の振る舞いが変化する境界は 1) $P = 0$, 2) $Q^2 = R^2$, 3) $P^2 = Q^2 - R^2$ の 3 つの曲線である。それぞれの式は k_1, αk_2, ω_2/ω_1 の関係によって表現され

1) より　　$\Omega = k_1 - 1$　　(3-8)

2) より　　$\alpha k_2 = (k_1 - 1 + \Omega)^2/4k_1\Omega$　　(3-9)

3) より　　$\alpha k_1 k_2 = k_1 - 1$　　(3-10)

Ω : 抑制層の興奮層に対する動作速度の比($=\omega_2/\omega_1$)

となるので、Fig.3-4に示すように、k_1 - αk_2平面上にその領域を表現した。平衡点が安定であるのは中の振動収束および無振動収束の領域であるから、その条件は(3-8),(3-10)より、

$$\Omega > \Omega_c = k_1 - 1 \qquad (3\text{-}11)$$

$$\alpha k_1 k_2 > k_1 - 1 \qquad (3\text{-}12)$$

ここで、　Ω_c : 臨界動作速度

である。(3-11)式より動作速度を速くした方が安定な平衡点が作り易い事がわかる。また、(3-12)式より αk_2の値を大きくしてゆくと安定な場合でも不安定な場合でも必ず振動的となることがわかる。

3-3-3. 平衡点の数およびその偶奇性と安定性の関係

微分方程式,の一般的な場合の振る舞いについて調べるために、平衡状態の方程式を導くと

$$x = F_1(x - \alpha y) \qquad (3\text{-}13)$$

$$y = F_2(x) \qquad (3\text{-}14)$$

となる。

これを x-αy 平面上で表現するとFig.3-5のようになる。ここで関数 $F_1()$, $F_2()$は値域 [0, 1]　の単調増加関数である。すると、(3-14)式は必ず単調増加関数であるが、(3-13)式は自己帰還のために多くの場合S字型にくねる。これより平衡点の数は1, 3, 5個の何れかである。ここで、x の小さい方から数えて奇数個目の平衡点を奇平衡点、偶数個目の点を偶平衡点と名づける。

次に平衡点の安定性を調べるために、平衡状態の方程式,式から y を消去して

$$G(x) = F_1[x - \alpha F_2(x)] - x \qquad (3\text{-}15)$$

を得る。そして $G(x)$ を x について微分すると、

$$\frac{dG(x)}{dx} = \left[\frac{dF_1(\xi)}{d\xi}\right]_{\xi = x - \alpha F_2(x)} \left[1 - \alpha\frac{dF_2(x)}{dx}\right] - 1 \qquad (3\text{-}16)$$

ここで、(3-4)式の定義より、

$$\left[\frac{dF_1(\xi)}{d\xi}\right]_{\xi = x - \alpha F_2(x)} = k_1, \qquad \frac{dF_2(x)}{dx} = k_2$$

であるから、

$$\frac{dG(x)}{dx} = k_1(1 - \alpha k_2) - 1 \tag{3-17}$$

するとに示すように、奇平衡点での先に示したG(x)の微係数は負で、偶平衡点では正である。つまり、

奇平衡点　　$dG(x)/dx < 0$　　$\alpha k_1 k_2 > k_1 - 1$

偶平衡点　　$dG(x)/dx > 0$　　$\alpha k_1 k_2 < k_1 - 1$

となる。よって収束の条件(3-12)式より、偶平衡点は必ず不安定となり、奇平衡点の安定性は式(3-11)次第である。つまり、安定な平衡点は奇平衡点であると同時に抑制層の動作速度が式を満たす程度に速いものである。

また、定常解の条件

$G(x) = 0 \quad (0 < x < 1)$ (3-18)

および

$$-x < G(x) < 1 - x$$
$$\lim_{x \to +\infty} G(x) = 1 - x, \qquad \lim_{x \to -\infty} G(x) = -x \tag{3-19),(3-20}$$

からも、Fig.3-6に示すように定常解の数が奇数個になる。

3-3-4. 計算機シミュレーション

まず関数$F_1(x)$, $F_2(x)$を線形関数と置いて平衡点近傍での解の振る舞いを調べたところ、収束性および振動性が理論と一致した。

次により実際的な状況に対するシミュレーションとして定性的にネットワークの活動度変換関数の特徴を反映しているシグモイド関数を用いた。つまり活動度変換関数を

$$F_i(\xi) = [\,1 + \exp(-a_i(\xi - \theta_i))\,]^{-1} \tag{3-21}$$
$$(i = 1, 2)$$

とした。その結果xが0または1に近い領域で現れる平衡点は$k_1 - 1$が負となるので(3-11)式より抑制層の動作速度がいくら遅くても安定になる。これに対し、xの値が中間的な値を取る平衡点では$k_1 - 1$が正の値をとるので、安定化のためには抑制層の動作速度Ω_cより

大きくする必要がある。

これらの知見に基づいて設定された適当なパラメータによる例をFig.3-7に示す。図中3つの平衡点のうち両端の2つが安定である。つまりxがほとんど0の領域に静止安定点が存在し、x が中間的な値（この例では約 0.3）の領域に活動安定点が存在する。また、活動安定点における抑制層の活動度は大きすぎず、活動点に収束する入力の範囲（情報レンジ）も適当である。なお、中の活動安定点は一つの点でしかないが、これは出力を表現するパターン空間では広がりを持ち、多くのパターンに対応している。

3-4. 興奮層をユニット化した研究

3-4-1. ユニット化した場合の力学

次の段階として出力レベル密度関数の制御を行う。そのため興奮層を複数のユニットに置き換えた。このモデルでは抑制層は興奮層の各ユニットからほぼ平均的に入力を受け、ほぼ平均的に出力を返す多数のユニットにより構成されている。またここでのシミュレーションでは興奮層の動作に興味がある。よって抑制層のユニットは興奮層の活動度に応じて出力を行う単一のユニットとして記述できる。

よって、システムは(3-1),(3-2)と類似の次の微分方程式に従うものとする。

$$\frac{dx_i}{dt} = \frac{1}{T_1}\left[f_1\left(\sum_{j=1}^{N} W_{ij}x_j - ay \right) - x_i \right]$$

$$\frac{dy}{dt} = \frac{1}{T_2}\left[F_2(x) - y \right]$$

$$\sum_{j=1}^{N} W_{ij} = 1$$

$$y = \frac{1}{N}\sum_{j=1}^{N} x_j$$

(3-22),(3-23),(3-24),(3-25)

x_i, y : 各ユニットの出力,

α : 抑制層から興奮層への結合係数,

T_1, T_2 : 興奮層と抑制層の時定数（$\omega_i T_i = 2\pi$）,

$f_1(), f_2()$: 各ユニットの動作関数

（シミュレーションでは、$f_i(\xi) = [1 + \exp(-a_i(\xi - \theta_i))]^{-1}$）

W_{ij} : 興奮層内の結合係数,

N : 興奮層内のユニットの数

また、においては興奮層内の結合の寄与は１に規格化していたので、(3-24)式に示すようにここでもその条件を踏襲する必要がある。この規格化が結合に対する唯一の制限であるが、実際にはある程度のばらつきが許容できると思われる。

3-4-2. 自己帰還入力の導入の動機

以上説明した条件の基でシミュレーションを行うと、活動度に関してはこれまでと同様

に良好な結果を得られる。しかし出力レベル密度関数に注目するとこのシステムは興奮層のユニットの出力が平均化をおこし、システムの表現するパターンが不明確となり、3-1.で示された目標は全く達成されない。

そこで目標とする出力レベル密度関数を実現するために、各興奮層のユニット毎に自分自身の出力を正帰還させる自己帰還入力を導入した。これにより上記(3-22)式を次の式に置き換えた。

$$\frac{dx_i}{dt} = \frac{1}{T_1}\left[f_1\left(\sum_{j=1}^{N} W_{ij}x_j + s(x_i) - ay \right) - x_i \right] \tag{3-26}$$

s() :　　興奮層各ユニットの自己帰還入力

これにより、一度出力の大きくなったユニットはますますその出力を大きくしようとするのでパターンが明確に存在する状態が安定となることが期待される。

3-4-3. 線形自己帰還入力

初期の試みでは自己帰還入力を線形とした。つまり、

$$s(x) = W_{SFI}\,x \qquad (W_{SFI} = \text{const}) \tag{3-27}$$

とした。

なお、シミュレーションでは各ユニットの動作関数は指数関数

$$f_i(\xi) = \left[1 + \exp(-a_i(\xi - \theta_i))\right]^{-1} \tag{3-28}$$

$$(i = 1, 2)$$

とし、興奮層内の結合はランダムに決定した。

以下にこのモデルでで比較的良好に動作した場合の条件をTable 3-1に示した。

		興奮層	抑制層
活動度変換関数の変数	指数関数の係数	$a_1 = 20.0$	$a_2 = 50.0$
	しきい値	$\theta_1 = 0.22$	$\theta_2 = 0.20$
動作速度		$\omega_1 = 1.0$	$\omega_2 = 3.0$
抑制層から興奮層への入力の利得		$a = 2.0$	
時間の刻み		$\Delta t = 0.001$	
ユニットの数		NUNIT = 10	
自己帰還の定数		$W_{SFI} = 0.5$	

Table 3-1 シミュレーションに用いたパラメータ

この条件の基では活動安定点と静止安定点が存在し、活動安定点での抑制層の活動度は興奮層のそれよりも小さい。活動安定点に収束する範囲（情報レンジ: IR）はある例では、

$$IR \approx 0.08 \ (x = 0.09 \sim 0.17)$$

であった。この値はネットワーク内の結合や初期入力のパターンに依存するので情報レンジが -0.02 ~+0.04 程度は変動する。また情報レンジ内のランダム入力を初期値とすると、ある程度の時間が経過した後に落ちつくパターンは２つのユニットがほぼ１で他のユニットがほぼ０である２種のパターンの何れかである。

このように自己帰還の比例定数(W_{SFI})を 0.5 とすると、明確なパターンを安定化する事ができる。しかしこの線形自己帰還のモデルではユニット自身の出力が大きくなる事により益々入力が大きくなるので、ある程度以上大きなの出力状態にあるユニットは自己帰還入力のみでも出力がほぼ１の状態に引き込まれる。同時にその反動で劣勢な出力状態にあるユニットの出力が強く抑制される（Fig.3-8 参照）。すると、その時現れたパターンが極端に安定化され、準安定な状態間を遷移する可能性が失われてしまう不都合がある。

3-4-4. 二次関数の自己帰還入力

上記の不都合を回避するために比例定数(W_{SFI})を 0.2 程度にすると状態の極端な安定化は回避できる反面、明確なパターンを表す状態に到達する事が困難となる。この状況はFig.3-8のようにユニットに対する入力(x_i)に対して入力和（ $\Sigma W_{ij}x_j + s(x_i)$ ）と興奮性ユニットの動作関数（ $f_1(x_i)$ ）を表現すると判りやすい。自発発振点は入力和と動作関数の交点であるから、線形で自発発振点が存在する場合にはFig.3-8に見られるようにその出力

はほとんど１である。

そこである程度以上大きい出力に対しては自己帰還入力を減少または飽和させればFig.3-9のように自発発振点が形成されてもその出力は直ちにほぼ１とはならず入力の大きさに応じて出力が変化するので、現れているパターンは適度に安定である。ここでは自己帰還関数の例として二次関数を用い、

$$s(\xi) = x_p(\xi/y_p)(2 - (\xi/y_p)) \qquad (3\text{-}29)$$
$$s(0) = 0, \qquad s(y_p) = x_p$$

とした。

以下にこのモデルでで比較的良好に動作した場合の条件を示すと、

		興奮層	抑制層
活動度変換関数の変数	指数関数の係数	$a_1 = 20.0$	$a_2 = 50.0$
	しきい値	$\theta_1 = 0.25$	$\theta_2 = 0.20$
動作速度		$\omega_1 = 1.0$	$\omega_2 = 3.0$
抑制層から興奮層への入力の利得		$a = 2.0$	
時間の刻み		$\Delta t = 0.001$	
ユニットの数		NUNIT = 20	
自己帰還二次関数の頂点座標		$(x_p, y_p) = (0.2, 0.5)$	

Table 3-2 シミュレーションに用いたパラメータ

である。この条件の基では活動安定点と静止安定点が存在し、活動安定点での抑制層の活動度は興奮層のそれよりも小さい。活動安定点に収束する範囲（情報レンジ: IR）はある例では、

$$IR \approx 0.08 \ (x = 0.07 \sim 0.14)$$

であった。しかしこの場合も前記と同様に情報レンジはある程度変動する

この条件のもとでの典型的な状態の変化をFig.3-10に示す。また、初期状態に情報レンジ内のランダム入力を与えると、ある程度の時間が経過した後に落ちつくパターンは比較的現れ易い似た様なパターンが数種類あるが初期値に依存してその様子は微妙に異なる。また、長時間シミュレーションを続行するとある場合には５つの準安定状態を次々に遷移する過程も観測された。

この様に自己帰還入力の関数を適切に選択する事により現れるパターンは適度に安定化する。つまり自己帰還入力の関数を適当に設定することにより活動時の出力レベル密度関数を少なくともある程度制御できる事が示された。

3-4-5. 過大な抑制出力に対する対策

これまでのモデルでは抑制層による抑制は興奮層に対して平均的に効果を及ぼす。しかし興奮層内の結合の偏差が大きい場合には強い結合が存在しそれら強い結合で結ばれたユニット郡が非常に強力な発振を起こすので、抑制層の抑制能力が追随できず抑制層の出力は過大となる。

そこで、抑制層の出力が正常の範囲のうちは興奮層のユニットに平均的に抑制をかけるが、その出力が過大になった場合には出力の大きな興奮性ユニットに対し選択的に強い抑制をかける機構を導入した。つまり従来の興奮層への入力が抑制層の活動度(y)に比例する αy であったのに対しここでは

$$\alpha (y + C y^2 x_i) \tag{3-30}$$

y：抑制層の活動度　　x_i：興奮層のユニットの出力　　C：定数

とした。この改良により強い結合が存在する場合の抑制層の活動度の増大を防ぐ事ができると同時に、他の良好な性質は保たれた。

3-5. 議論

3-5-1. 過大な入力の排除と情報レンジ

3-5-1-1. 情報レンジと動作速度比

我々が考えているシステムでは、意味を持つ情報パターンは付加的機構により制御された状態と仮定しているので、過大な入力は情報的意味を持たない。

過大な入力があり一旦興奮層の活動度 x が大きくなった場合はネットワーク内のユニットに対する抑制入力の比率がネットワーク内の情報に対して大きくなりすぎるので、本来の処理に悪影響を与え不都合である、特にほとんど全てのユニットが興奮するような大入力ではネットワークは最も安定ないくつかのモードに到達するのみで情報処理装置として機能しない。よってシステムは過大な入力には反応しない方がよい。本論文におけるモデルはこの機能を実現することが可能であることを以下に示す。

抑制層の動作速度を大きくすると活動安定点の引き込み力は強くなり、広い範囲の入力が活動安定点に収束する。そこで、過大な入力があった際には活動安定点に収束しないように抑制層の動作速度を適度に遅くした方がよい。

Fig.3-7活動度によるシミュレーションを用いて静止しているネットワークに対して瞬間的に様々の大きさの入力を与えた（初期値が (x, 0) ）場合の活動安定点に引き込まれる x の範囲である情報レンジについて興奮層と抑制層の動作速度の比($\Omega = \omega 2/\omega 1$)を変化させて調べた。

に示すように、x-αy 平面上の x の小さい側からそれぞれn番目の平衡点の座標を (x_n, αy_n) (n = 1,2,3,4,5)とする。

入力活動度がが第２平衡点よりも小さい場合($x < x_2$)には第１平衡点に収束するので、情報レンジの最小値は x_2 である。一方最大値は動作速度比(Ω)に依存するので、その値を臨界最大活動度を x_c とする。すると情報レンジは、

$$IR = x_c - x_2 \tag{3-31}$$

である。

シミュレーションより、臨界最大活動度(x_c)は

$$x_3 < x_c \leqq 1 \tag{3-32}$$

を満たし、動作速度比(Ω)がある程度以上大きいと1になり、反対に前記(3-11)式による収束条件をどうにか満たす程度の場合にはほとんど x_3 である。

よって、情報レンジを適切な大きさにするためには動作速度比が臨界動作速度比よりも僅かに速い程度

$$\Omega \geqq \Omega_c \tag{3-33}$$

に、設定すればよい。

3-5-1-2. 興奮層の活動度変換関数の学習による変化

シミュレーションでは活動度を平均化したモデルを用いたが、実際には出力パターン毎に活動度変換関数は異る。活動度変換関数のばらつきが少ない場合にはあらゆる出力パターンに対して適切な諸条件を設定できるが、そのばらつきが大きくなるといくつかのパターンが安定でなくなる一方あるパターンは情報レンジが必要以上に拡大する問題が起きる。

外界に依存する学習がなされていない初期の段階では、ネットワーク内の結合は等方的で活動度変換関数のばらつきが小さいのでネットワークは適切な条件を実現できるが、学習が進行すると活動度変換関数のばらつきが大きくなり、上記のような問題が発生する。この問題を避ける方法として動作速度比(Ω)を大きくすることにより不安定になった出力パターンを安定化し記憶の消失を防ぐことはできるが、何れにしろ情報レンジの拡大は不可避である。上記の回避方法を用いる際に抑制層の動作速度を速めることができない場合（おそらく生物系ではそうであろうが）には動作速度比(Ω)を大きくするために興奮層の動作速度を遅くしなければならない。この様な場合このシステムでは学習の進行とともに1).動作速度が遅くなる、2). 記憶された出力パターンの出現頻度が片寄る、等の状況が発生すると予測される。

3-5-1-3. 情報レンジの微小入力側への拡大

ネットワークの学習により生じる活動度変換関数のばらつきは情報レンジを微小入力側に拡大する効果も生むであろう。ある入力信号パターンが繰り返し提示されると、その入力信号と誘発される出力パターンを結び付けるネットワーク上の結合が増強される。するとその入力信号はネットワーク対し通常よりも大きな効果を与える。この様な場合入力信号の情報レンジは x_2 を越えて微小入力側へ拡大しえる。

3-5-2. 活動安定点における無振動収束解の存在可能性

静止安定点の収束の形態は一般に無振動的であるのに対して活動安定点は一般に振動的な収束形態を示す。しかし、振動的な振る舞いは一時的な抑制層の活動度の増大による処理情報のダメージを引き起こすので情報処理上は好ましくない。

3-3.での研究より αk_2 の値を小さく保てば解は無振動な安定点になることが判ってい

る。無振動的となる十分条件は(3-9)式のk_1をパラメータとした最小値より

$$\alpha\ k_2 < (\Omega-1)/\Omega \tag{3-34}$$

である。

一方安定点であるための条件はより、x-αy平面上において抑制層に関する平衡曲線の傾きが興奮層の平衡曲線の傾き(L)よりも大きいことであるから、

$$\alpha\ k_2 > L \tag{3-35}$$

という条件を満たす必要がある。

静止安定点では興奮層の平衡曲線の傾き(L)が小さいのでαk_2 を小さい値に保つのは容易であるが、活動安定点では傾き(L)は抑制層の分散が小さくなるに伴い１に近づくので上記二つの条件を満たすαk_2 をさがすのは非常に難しく、我々が行ったシグモイド関数を用いたシミュレーションではこれまでのところ発見できない。つまり活動安定点は通常振動的であると考えるべきであろう。

さらに、前章で述べたような活動度変換関数のばらつき等の影響を考慮すると、活動安定点において,の２つの条件を成立させ続けることはほぼ不可能である。またこの条件を満たすことができたとしても興奮層と抑制層の平衡曲線の傾きが非常に近づくので、興奮層の特性が僅かの変動が活動安定点のｘの値に大きく影響し記憶されたパターンが変化を受ける問題が生ずる。つまり無理に無振動的な活動安定点を作るのはあまり有益でない。

3-5-3. システム内のパラメータの性質

3-3-1.で説明したこのモデルの状態を実現するために、主な７つのパラメータをどのように制御すれば良いか検討した。シミュレーションでは活動度変換関数は式と同様のシグモド関数とおいたので、しきい値と分散と強い関係を持つ係数(ai)について検討した。議論のポイントはは制御目標が存在するか、制御が可能であるか、学習によりどの様な影響を受けるか等である。

1). 興奮層内帰還結合の規格化(モデル中では１)

で説明したように興奮層内のフィードバック入力を規格化するために、

$$\sum_{j=1}^{N} W_{ij} = 1 \tag{3-24}$$

を満たす必要がある。そこで、学習処理毎に、

$$W_{ij} \rightarrow \frac{W_{ij}}{1 + \sum_{j=1}^{N} \Delta W_{ij}} \qquad (3\text{-}36)$$

ΔW_{ij} :　学習則による結合の変化量

というように規格化を行う。よって、各ユニットはこの処理を独立に実行することができる。

2). 興奮層の活動度変換関数の分散(1/a_1にほぼ比例)

興奮層の活動度変換関数の分散は学習を通して外界と相互作用するネットワーク内の結合により決定されるので、自由に制御する事はできない。

3). 興奮層の活動度変換関数のしきい値(θ_1)

興奮層のしきい値は興奮性の各ユニットのしきい値を一定にすれば、活動度変換関数のしきい値もその値となる。そこで各ユニットのしきい値を同時に変化させれる事により活動度変換関数のしきい値は制御できる。

4). 抑制層の興奮層からの入力の規格化(モデル中では１)

抑制層への入力の大きさは抑制層の活動度変換関数の分散としきい値との比較する事により意味を持が、以下では簡単のために１に規格化しておく。

規格化は上記　1)と同様に入力の結合の和を一定に保つ事によりなされるが、抑制層は外界の影響による学習を行わないので一度ネットワークが形成されればそれ以降はあまり制御の必要はない。この場合も処理は抑制層内のユニット毎に実行できる。

5). 抑制層の活動度変換関数の分散(1/a_2にほぼ比例)

抑制層の分散は興奮層の活動時の活動度を固定する精度に関する変数であり、ある程度以下に小さければよい。

4)の規格化の精度と次に述べる6)のしきい値制御の精度をあげれば分散を小さくできる。また、抑制層のユニット同士で同時に興奮するよう互いにしきい値を調整すれば、分散をかなり小さくできるだろう。

6). 抑制層の活動度変換関数のしきい値(θ_2)

抑制層のしきい値は、興奮層の活動時の活動度を決定するパラメータである。全てのユニットのしきい値を同様に変化させれば制御できる。しかしこのパラメータも一度適切な値に設定すればそれ以上は制御する必要が無い。

7). 抑制層から興奮層への利得(α)

この値はある程度以上大きい値に一度設定すればそれで良い。

以上のような各パラメータの性質から、興奮層の分散２）は制御でない上に学習によって変化してしまい、興奮層のしきい値３）は制御目標が設定できない。しかし他のパラメータは制御可能で制御目標も存在する。今後の議論は制御できない興奮層の分散の変化に対していかにしてネットワークの状態に健全に保つかという点である。

3-5-4. 異常状態に対する考察

上記したパラメータが適切に制御されればシステムは健全な状態を保つことができるがしばしばこの状態を逸脱することが考えられる。異常状態の主な原因は学習により興奮層の活動度変換関数の分散が変化である。

3-5-4-1. 活動安定点の消失

活動安定点が消失した場合は記憶されている情報が取り出せなくなると同時に短期的な記憶が保持できないという傷害を引き起こす。この状態はネットワークが大きな知的システムの一部である場合には一部の情報を一時的に取り出せなくなるだけで、パラメータを適当に再調整する事ができればそれらの情報は再び利用できるので大きな問題ではない。人間で言えばど忘れの状態であろう。

3-5-4-2. 静止安定点の消失

静止安定点が消失した場合、ネットワークの状態は活動安定点にロックされ、ある出力パターンが恒常的に出力される。この場合はサブシステムであっても大きなシステム全体に影響を与える。人間では一つの考えにとりつかれたようにな症状であろう。生物系では疲労による出力の低下により永久的な状態のロックは避けられるが、もしこの状態が持続するとシステムの正常な機能は阻害されるだろう。

3-5-4-3. その他の異常状態

これ以外にもパラメータの不適切な設定が様々な異常状態を引き起こすた考えられるが、ここではサブシステムとしてこのネットワークを含む大きなシステムへ悪影響を与える異常状態について考える。すると興奮層の活動度がほぼ１の安定点が形成されてネットワー

クが完全に興奮し続ける状態が生じた場合が深刻である。この異常状態は抑制層から興奮層への利得が小さくなり、興奮層の活動度がほぼ1の領域に安定点が形成された場合に相当する。現在のところこれは人間におけるテンカンの対応する症状ではないかと推測している

3-5-5. パラメータの制御

活動度が半分以下の領域に活動安定点を設定した場合には、 Fig.3-11(a)に示すように、興奮層の分散が小さくなると興奮層の平衡曲線（x-αy平面上で記述される）が減少し、ある限界を越えると活動安定点が消失し、一方逆に分散が大きくなりすぎると静止安定点が消失する。この分散の変動に対して興奮層の活動度変換関数のしきい値をいかに制御するかがここでの課題である。

静止安定点が消失したネットワークはFig.3-11(b)に示すように活動状態にロックされるのでその出力の長時間平均は明らかに正常な場合と異なる。よって出力平均の増大を感知して興奮層の活動度変換関数のしきい値を増大させる事で対応できる。Fig.3-11(c)にみられるように、活動安定点が消失した場合には逆にしきい値を減少させる必要がある。だがこの場合におけるネットワークは興奮状態を保持することはできなくても外部からの入力によって過渡的興奮は起こるのでその消滅を感知することは難しい。最も簡単な方法は恒常的にしきい値を僅かづつ減少させる方法である。これにより静止安定点の消失が起こるが、前記の回復方法を併用すれば不都合はない。

上記の二つの機能がバランスすればネットワークを健全な状態に保つ事は可能である。

3-5-6. ユニットの数と安定状態の関係

シミュレーションにおいてユニットの数を少なくすると状態の制御がうまく行われずユニットの出力が平均化する事が観察されている。この機構を解明するために抑制層の高速近似を用いて考察を行った。

興奮層のユニットの微分方程式は、次のように記述できる（参考(3-22)式）

$$\frac{dx_i}{dt} = - x_i + f\left(x_i, \sum_{j=1}^{N} W_{ij}x_j, y, s_i\right) \qquad (3\text{-}37)$$

x_i : 興奮層のユニットの出力,　　t : 時間,

W_{ij} : 結合係数　y : 抑制層の活動度,　　s_i : 外部入力

これはつまり

$$\frac{dx_i}{dt} = F\left(x_i, \sum_{j=1}^{N} W_{ij}x_j, y, s_i\right) \tag{3-38}$$

と、記述できる。ただし、F は次の条件を満たす必要がある。

$$F\left(0, \sum_{j=1}^{N} W_{ij}x_j, y, s_i\right) > 0$$

$$F\left(1, \sum_{j=1}^{N} W_{ij}x_j, y, s_i\right) < 0 \tag{3-39),(3-40}$$

ここで、

$$0 \leqq x_i \leqq 1, \qquad 0 \leqq \Sigma W_{ij}\, x_j \leqq 1, \qquad 0 \leqq y \leqq 1$$

である。

ここでは、一つの興奮性ユニットの平衡状態について検討する。そこで、注目したユニットの抑制層を介しての自身への寄与を見積もるために、抑制層の動作速度が無限に速いと仮定した。これはx_iが平衡点近傍ではゆっくりと変化するために、相対的に抑制層の動作速度が非常に大きくなると考えられるためである。なお、注目した以外の興奮性ユニットの寄与によるyの値はいろいろと変化させる。この様に表現された関数を$F_{HL}()$と表現する。そこで、ネットワークの特性を調べるために、(3-38)式において $dx_i/dt = 0$ とすると、

$$F_{HL}\left(x_i, \sum_{j=1}^{N} W_{ij}x_j, y, s_i\right) = 0 \tag{3-41}$$

となる。

そして$\Sigma W_{ij}\, x_j$ に対する x_i の変化を表現すると、我々が望む曲線の形状はFig.3-12のような形である。なぜならこれまで述べてきたように、ユニットの出力がハイレベルとローレベルが明確に分裂し、かつ、それぞれのレベルにおいてある程度広がりを持って分布することが望まれると、同時にヒステリシスをなるべく持たないことが望まれるからである。ある x_i に対してひとつの $\Sigma W_{ij}\, x_j$ しか持たない場合にはそこがユニット出力の安定な値であるが、３つの $\Sigma W_{ij}\, x_j$ の値を持つ場合には双安定となる。この、双安定な場合がユニットの出力がヒステリシスを持っている状態である。そして曲線中において、$\Sigma W_{ij}\, x_j$ の極大値と極小値の差が大きいほど、ヒステリシスの特性が大きくなる。

これまで行ってきた結果から比較的良好に動作する条件で同様なグラフを描く。シミュレーションでは、関数F()としてシグモイド関数を利用したので、入力和は次の式で表される。

$$\sum_{j=1}^{N} W_{ij}x_j = -\frac{1}{a_1}\log\left(\frac{1 - x_i}{x_i}\right) - x_i(1 - x_i) + \alpha y^2 x_i + \theta_1 + \alpha y \qquad (3\text{-}42)$$

描いた結果Fig13(a)-(f)を見ると、ユニット数が大きく抑制層を介しての効果を無視できる場合には、その曲線の形状はほとんど我々が望んだ形となっている。しかし、ユニット数が少ない場合には、抑制層を介しての効果が相対的に大きくなるために、グラフの様子は我々が望む形から大きく離れてしまう。また、y^2の項の寄与によりx_iの大きな領域により強く抑制が働くことが示される。

これにより、いくつかのシミュレーションにおける出力が平均化してした状態やヒステリシスの機構が解明できた。

3-5-7. 抑制ユニットの動作速度と生体的知見との関係

生体に対する実験では興奮性ユニットの反応に引き続き抑制性ユニットの反応が起こる事が知られている。これに対し本論文では抑制性ユニットの動作速度が興奮性ユニットのそれよりもある程度速い方が好都合である事を主張した。この二つの主張は一見矛盾しているので、本論文におけるシステムは生体とは異なっていると考える事もできるが、考え方次第では必ずしも矛盾しない事を以下に示す。

この議論の対象として適切な入力の大きさの範囲は第２平衡点(x_2)より少し大きい程度の範囲である（Fig.3-7参照）。なぜならこれより小さい範囲では興奮性ユニットの出力が自然消滅してしまうので抑制性ユニットの反応が起きない。また初期の段階で興奮層の活動度が大きいと、引き続く興奮性ユニットの出力の増加が起きずに抑制性ユニットの出力が増大する。この過程では、初期の段階ですでに興奮性ユニットの反応が先行していると解釈できる。

第２平衡点より少し大きい範囲の入力が与えられた場合に２つの層の活動度の時間変化はのようになる。図中興奮層の活動度の増大する領域はＡであり、抑制層においてはＢである。よってこのシステムを観測すれば、興奮性ユニットの反応に引き続いて抑制性ユニットの反応が起こるので、前記２つの主張は矛盾しない。

3-5-8. 状態制御を含んだネットワークの動作方式

これまでのモデルでは生物学的知見に沿ったしきい値素子的なユニットにより構成されるネットワークに対してその活動状態を制御する機構を付加した。だが、工学的な視点に立てば予め状態制御の機能を含む動作方式を用いた方がより効率的で見通しの良いモデルとなるだろう。

ここでは本論文おける連続出力連続時間モデルを単純化し、離散出力離散時間としたモデルを提案する。抑制層の働きは活動時の興奮層の活動度を一定に保つ事、および静止安定点が存在する事を考慮すると、あるユニットの入力和がネットワーク内で大きい方から数えて決まった数以内で、しかも入力和がしきい値以上の場合に1を出力し他の場合に0を出力する、という動作方式を用いればよい。

以上のように単純化されたモデルはデジタルコンピューターを用いてシミュレーションを行う場合に非常に計算速度が向上する。しかしアナログモデルにおける多くの特徴を残しているので、大規模なシミュレーションを行うに際して有効な単純化となり得る。

3-6. まとめ

ニューラルネットワークの機能を向上させる方法として、その情報処理層の出力が実現し得る安定状態を制限する方法を提案した。これによりネットワーク内の結合に有効に情報を蓄積することができる。具体的には情報処理を行う興奮性ユニットよりなる相互結合型のネットワークに対して、抑制ユニットよりなるフィードフォアード型のネットワークを付加した。

一般にN個のユニットにより構成されるネットワークの状態はN次元超立方体中の任意の1点として記述でき、その変化は超立方体中の遷移ベクトルにより記述される。単層のネットワークでは超立方体中の遷移ベクトルはユニット間の結合のみにより記述される。しかし、通常ネットワークにはその使用目的に応じて様々な機能が要求されるので、ある種のネットワークではN次元超立方体中の全ての領域が有用ではない。そうした場合超立方体中の全ての領域における遷移ベクトルをユニット間の結合を用いて規定するのは効率が悪い。よって、他の機構により有用な領域のみに安定状態が存在するように閉じ込めを行い、結合は有用な状態間の遷移ベクトルのみを規定すればネットワークの能力は向上するであろう。

このシステムは他にもいくつかの特徴を持つ、1)　学習を行っていない段階でも尤もらしいパターンを表現している状態が安定で、冗長な隠れユニットにも出力が現れる、2) ユニットに対する入力の規格化をのぞいては結合がどの様な値をとっても必ず有用な領域に対応するパターンを表現するので、従来に比べ自由な学習則が設定できる、3)　情報処理層の結合が興奮性のみのモデルでしかも自己組織的に構成する事が可能なので、生体内で実現できる可能性がある、4) お互いに容易に遷移可能な自律安定状態を実現できる。

今後は実際に学習則を導入する事や、単純化したモデルで冗長な隠れユニットの振る舞を調べたり複数のネットワークを連結してその相互作用を調べる事が課題となる。

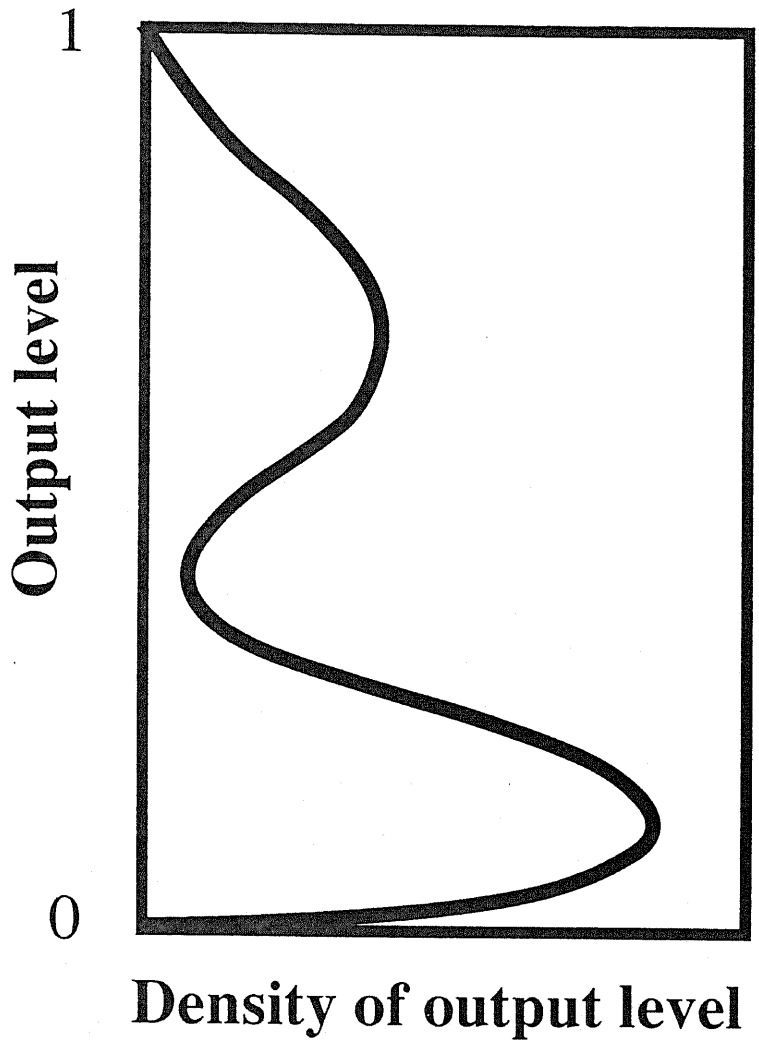


Fig. 3-1 Density of Output Level States Function of Fuzzy Bistable State

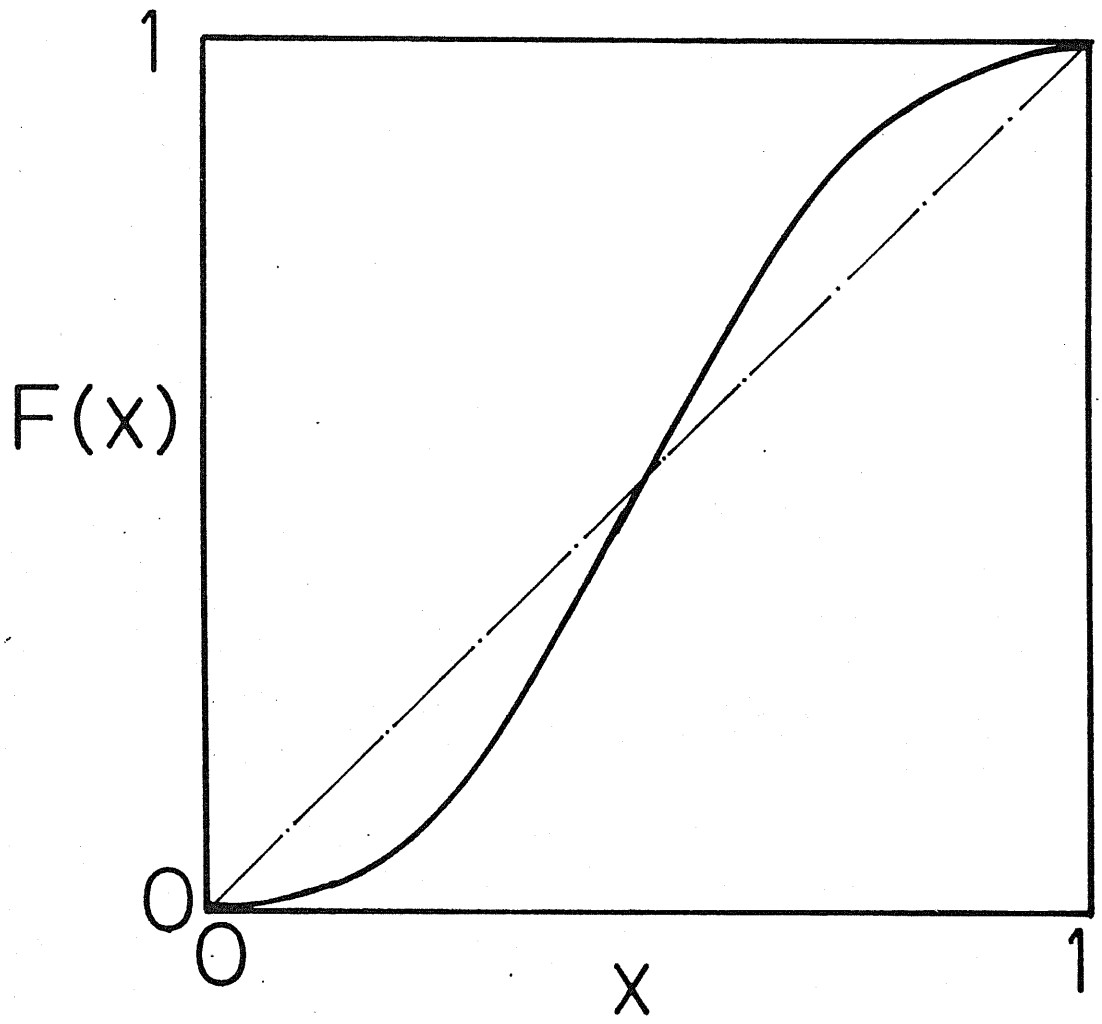


Fig.3-2単純な相互結合型ネットワークの
活動度変換関数

活動度X(t)の定義は、

$$X(t) = (1/N) \sum x_i(t)$$

N:ユニットの数，$x_i(t)$:ユニットの出力

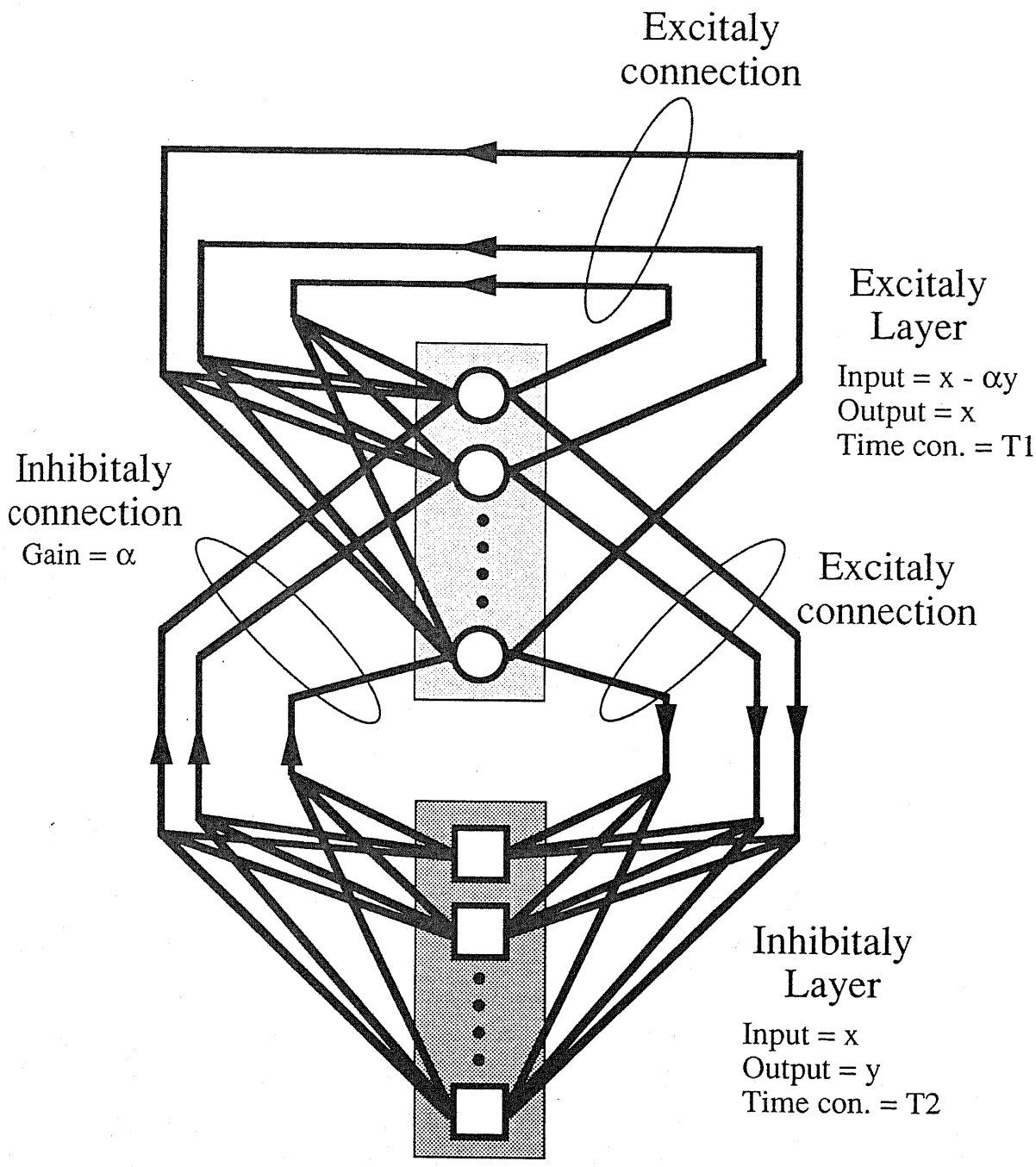


Fig.3-3 Tow Layers Model

○ Excitaly unit
□ Inhibitaly unit

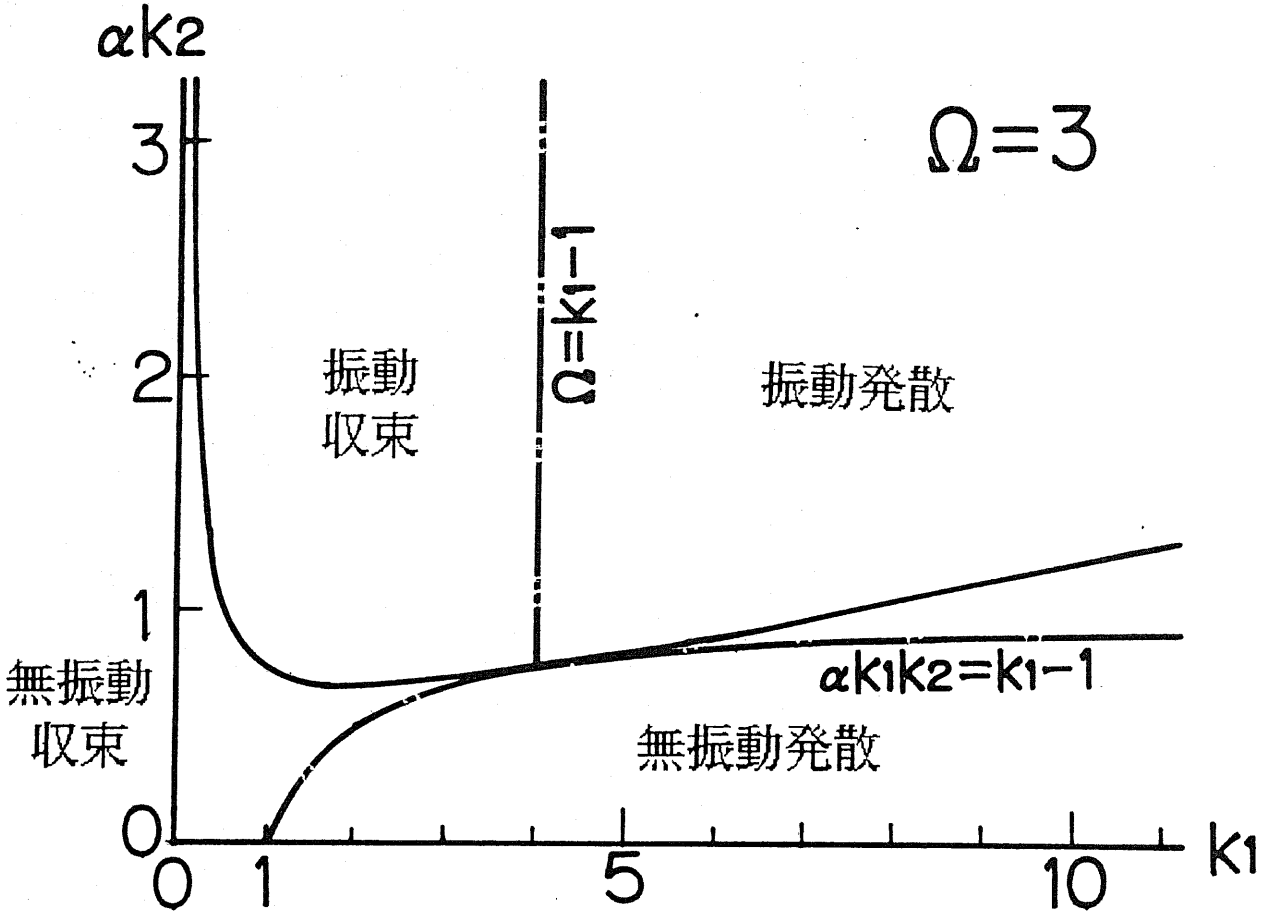


Fig.3-4 平衡点の性質($k_1-\alpha k_2$ plot)

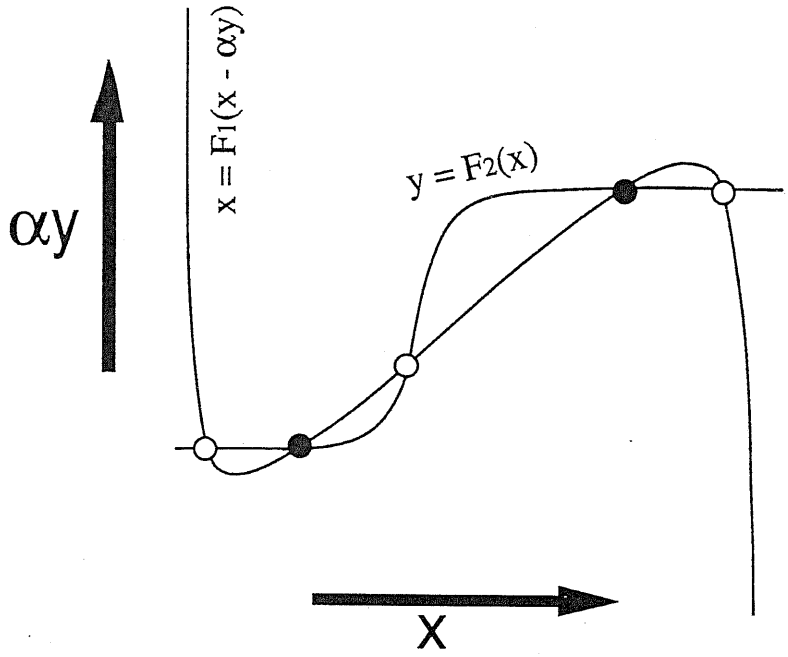


Fig. 3-5 Equilibrium Equation of Two Layer on x-αy Coordinate

○ stable equilibrium poiint

● unstable equilibrium poiint

ここでは、平衡点が五つある場合を示したが、一般には１、３、５個の何れかである

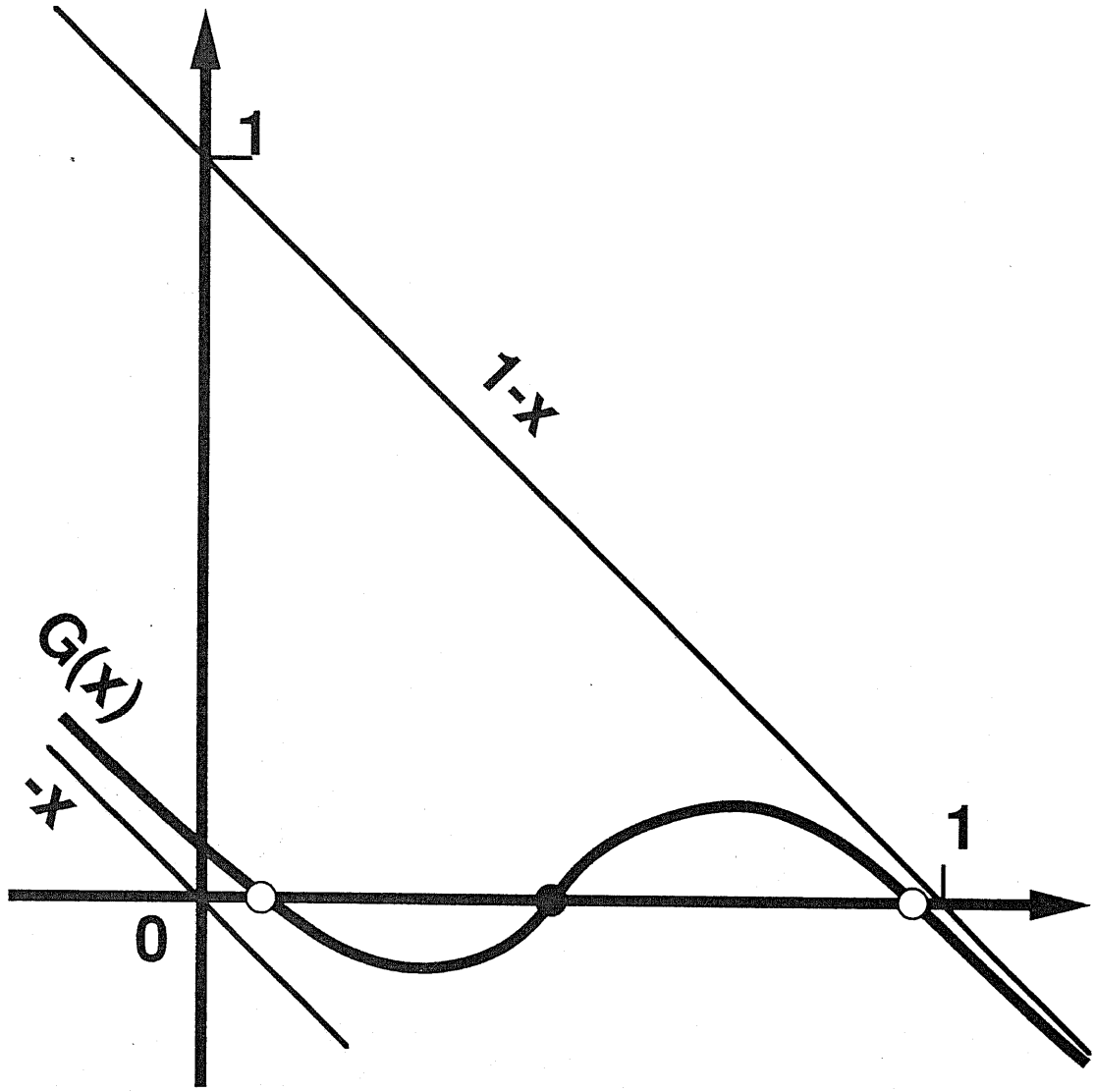


Fig. 3-6 Number of Equilibrium Points

○ stable equilibrium poiint

● unstable equilibrium poiint

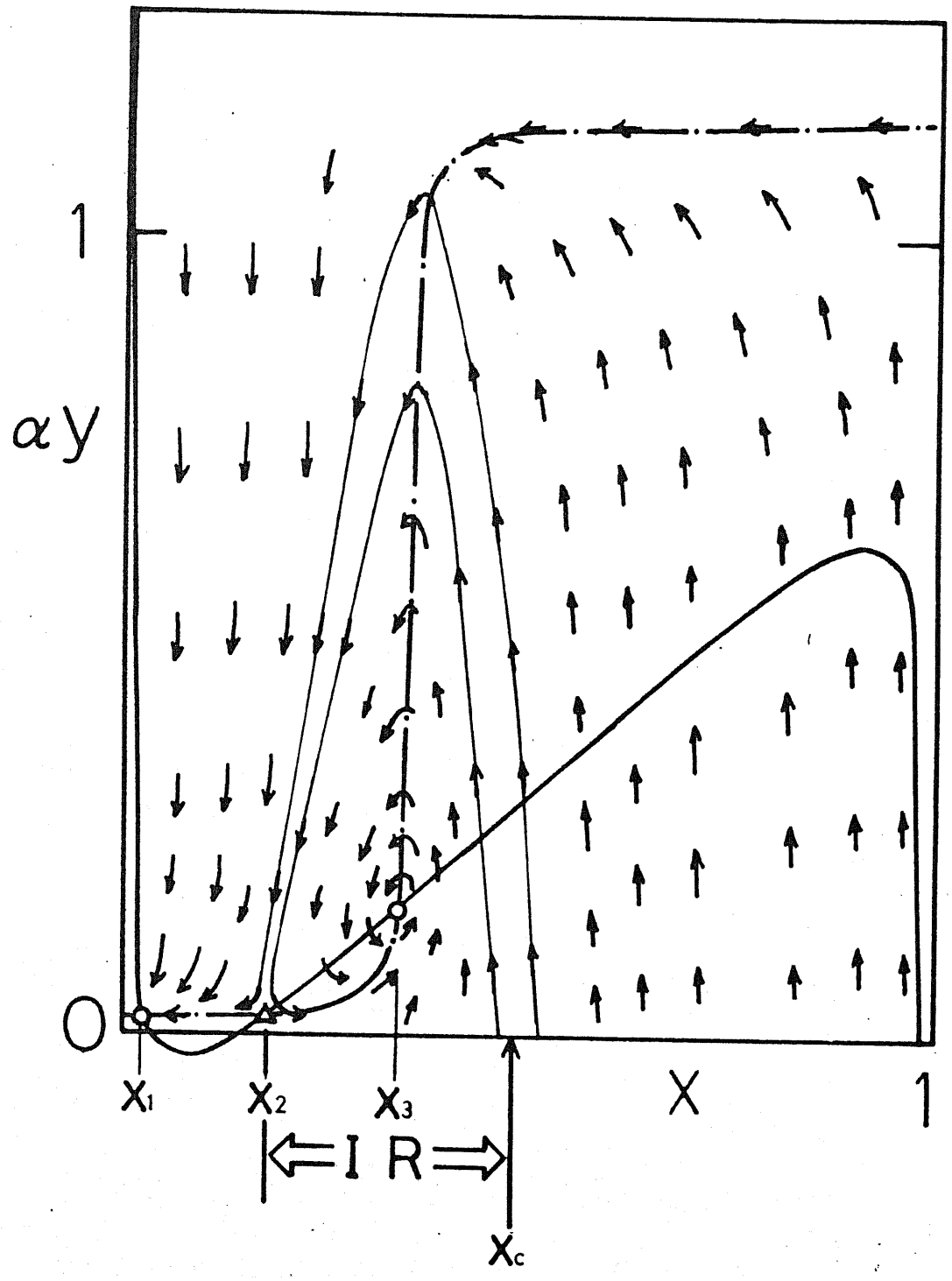


Fig.3-7活動度遷移図（シグモイド関数）
○…安定平衡点、　△…不安定平衡点
$\Omega \simeq k_1 - 1$　の時　$x_c \simeq x_3$
$\Omega >> k_1 - 1$　の時　$x_c = 1$

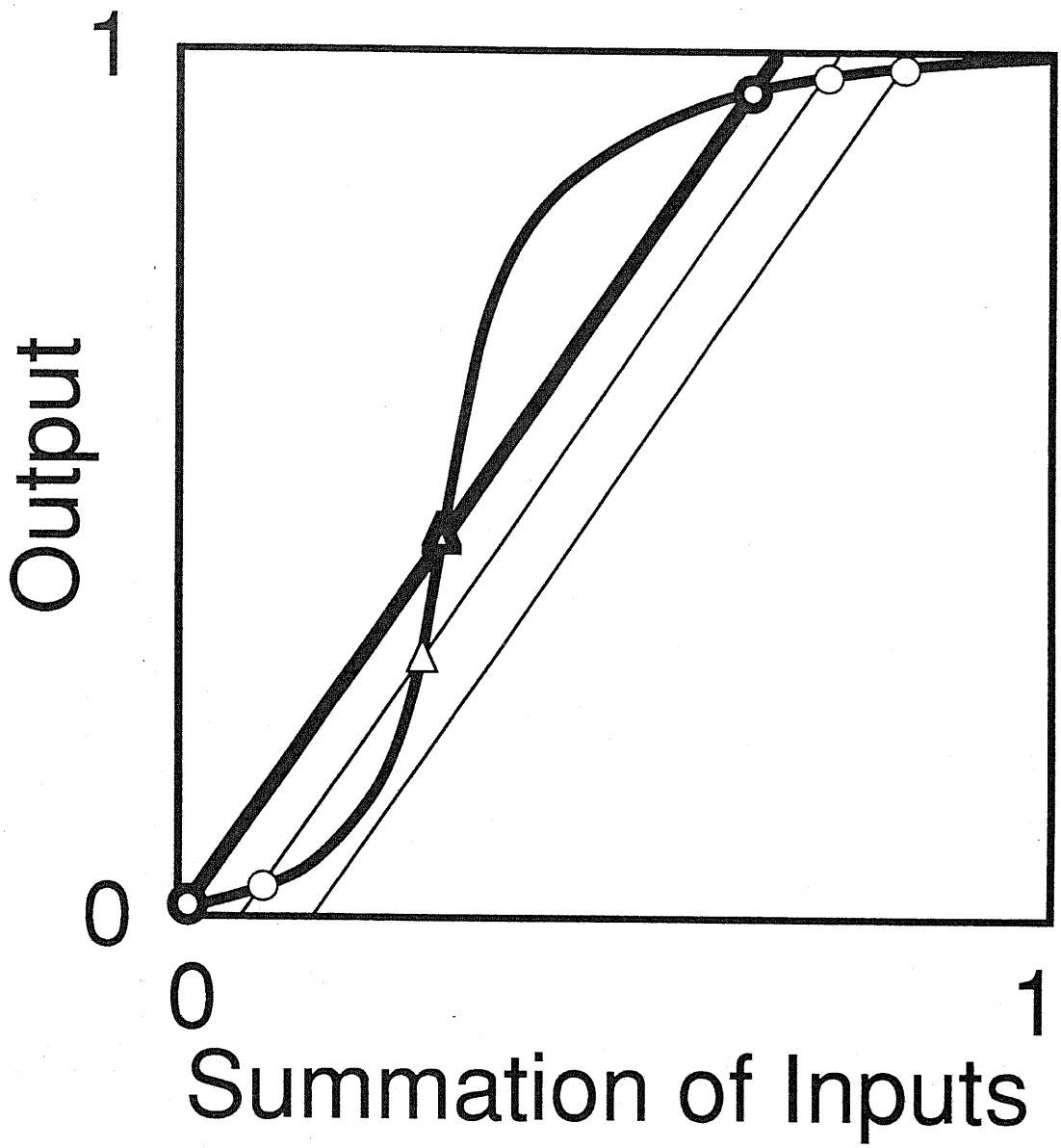


Fig. 3-8 興奮層ユニットの入出力関数および線形関数の自己帰還入力

● 自己発振を起こす安定平行点
▲ 不安定平行点

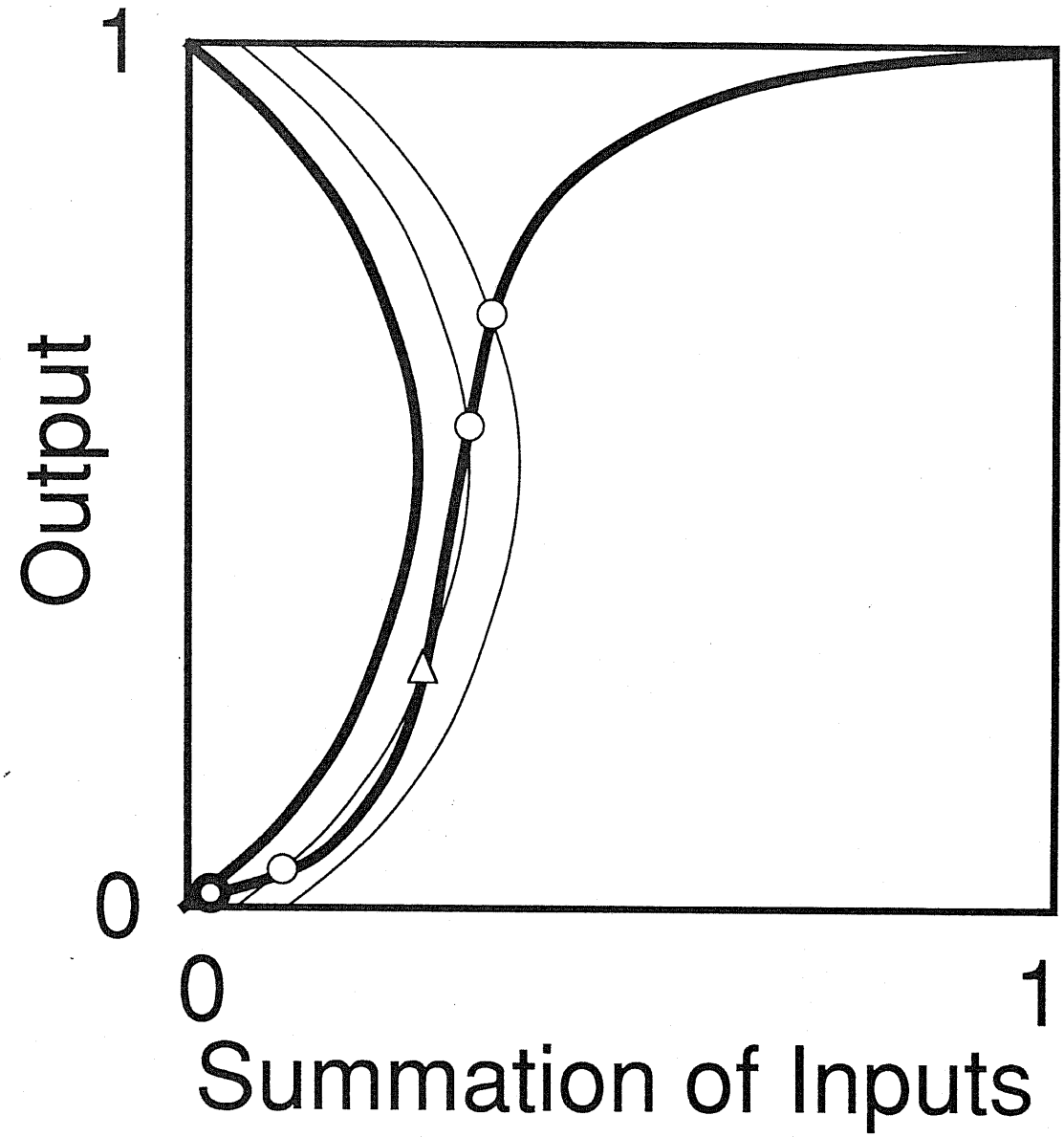


Fig.3-9 興奮層ユニットの入出力関数および二次関数の自己帰還入力

○ 自己発振を起こす安定平行点
▲ 不安定平行点

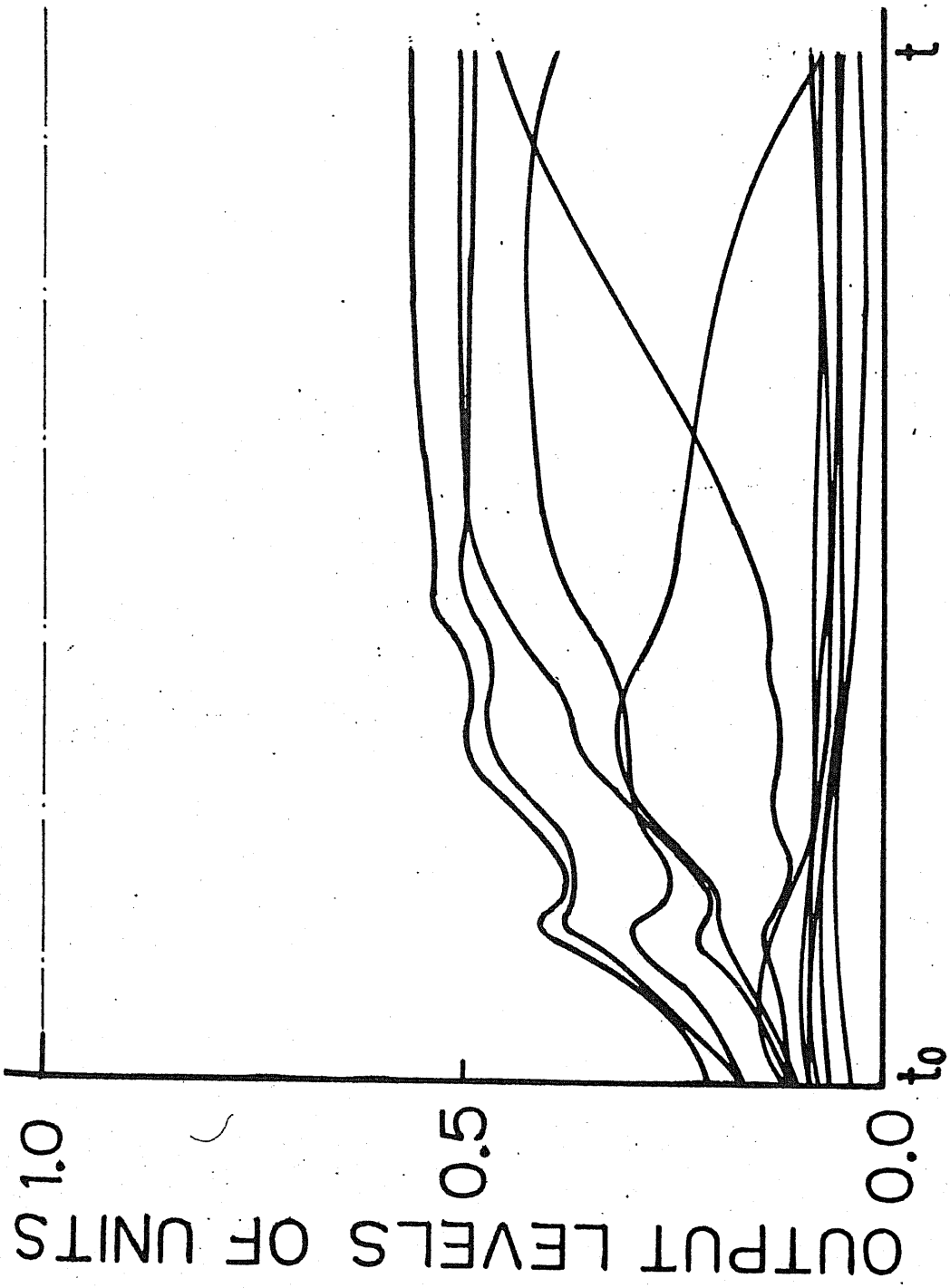
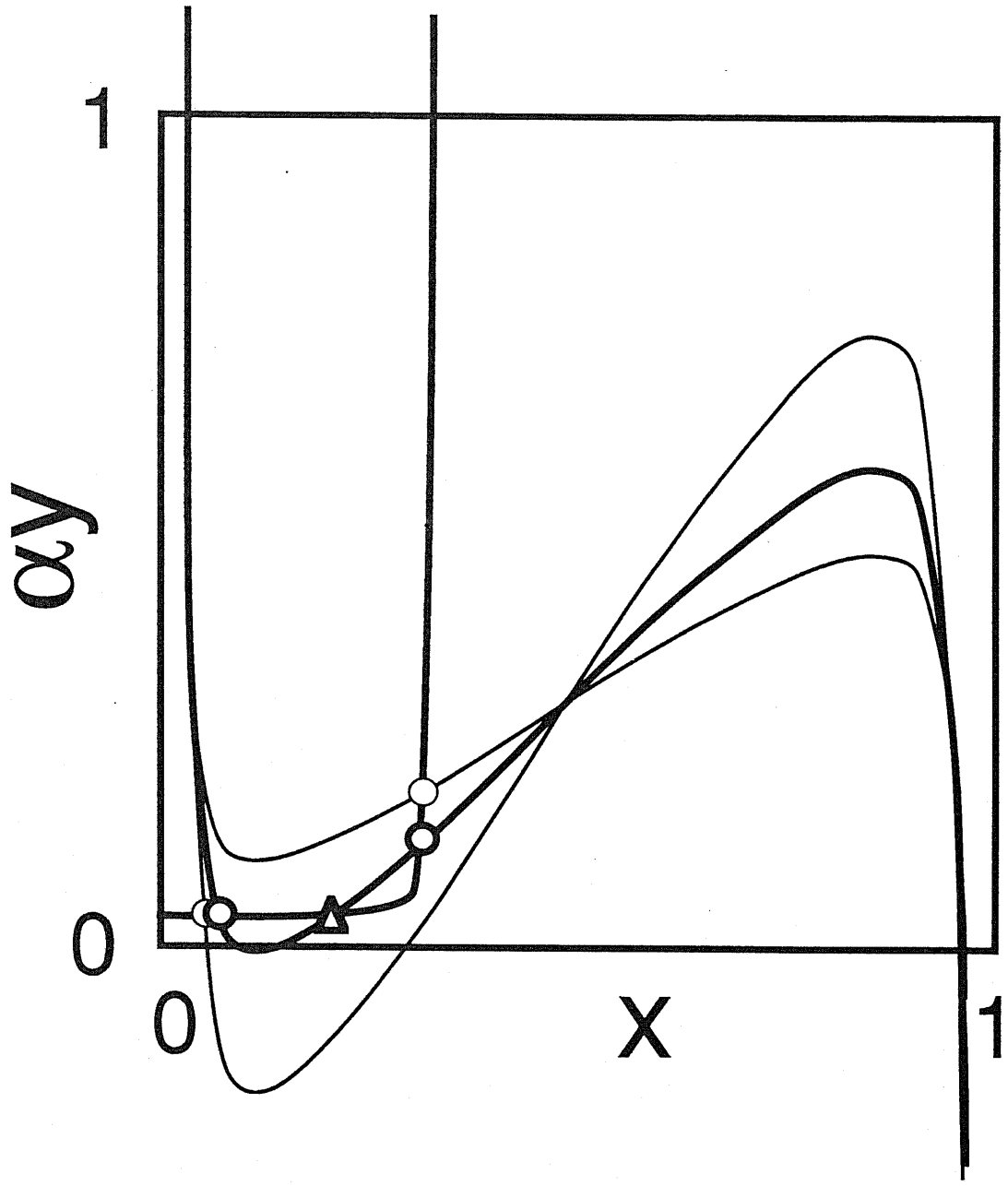
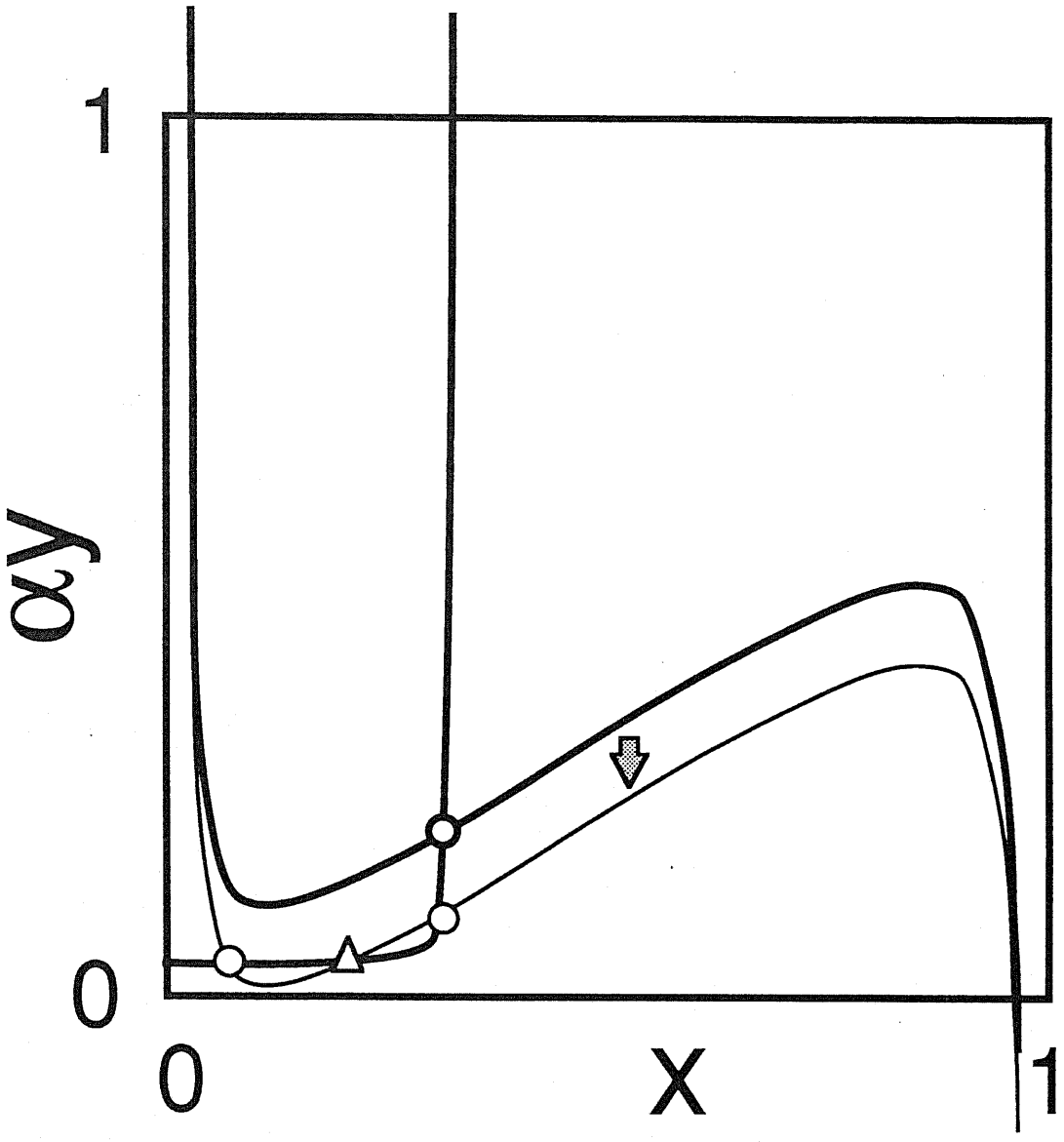


Fig.3-10 ユニット出力の経時変化
（二次関数の自己帰還入力を用いた場合）



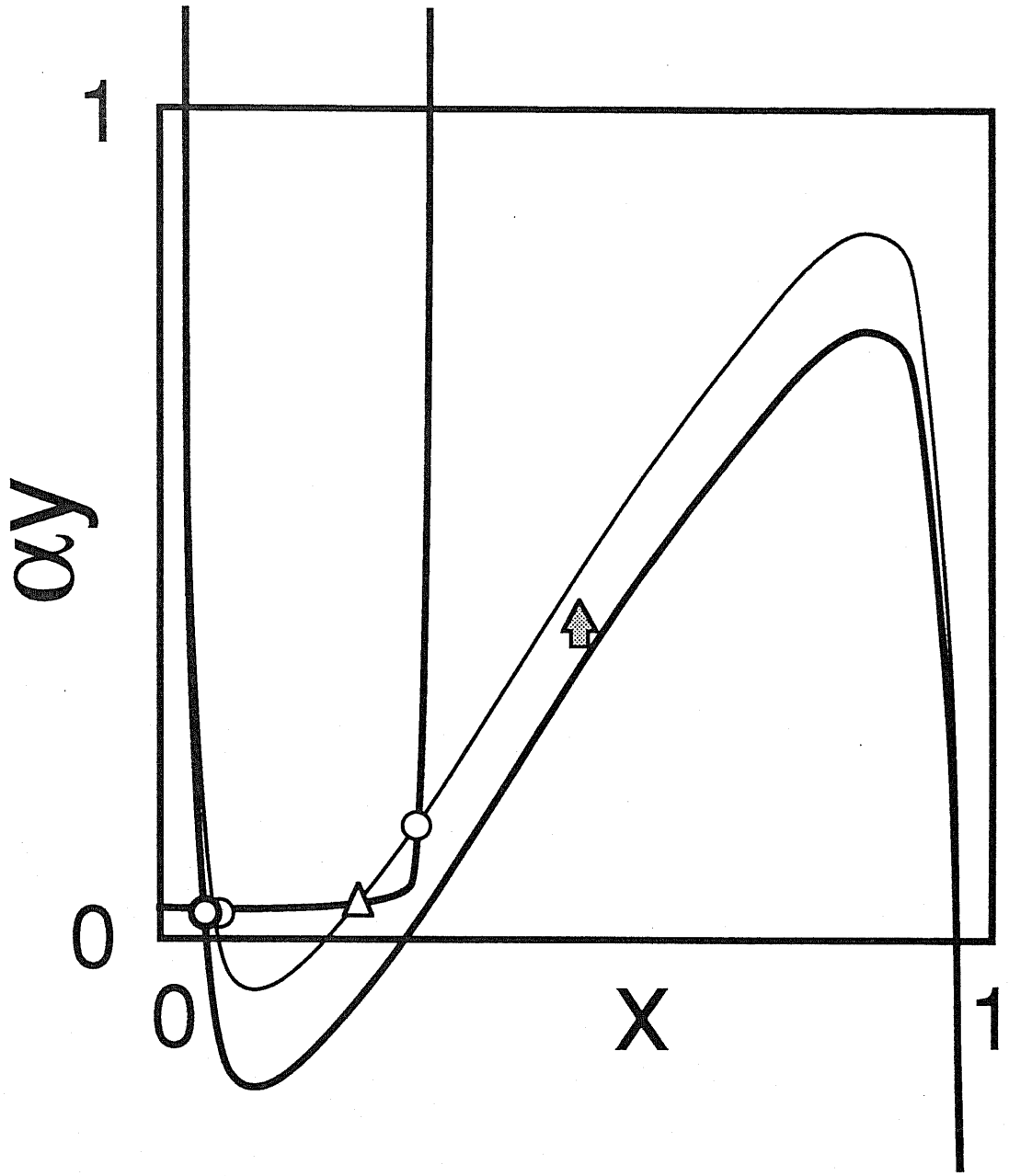
○ 自己発振を起こす安定平行点
△ 不安定平行点

Fig.3-11(a) 異常状態の発生



○ 自己発振を起こす安定平行点
△ 不安定平行点

Fig.3-11(b) 異常状態からの回復



○ 自己発振を起こす安定平行点
△ 不安定平行点

Fig.3-11(c) 異常状態からの回復

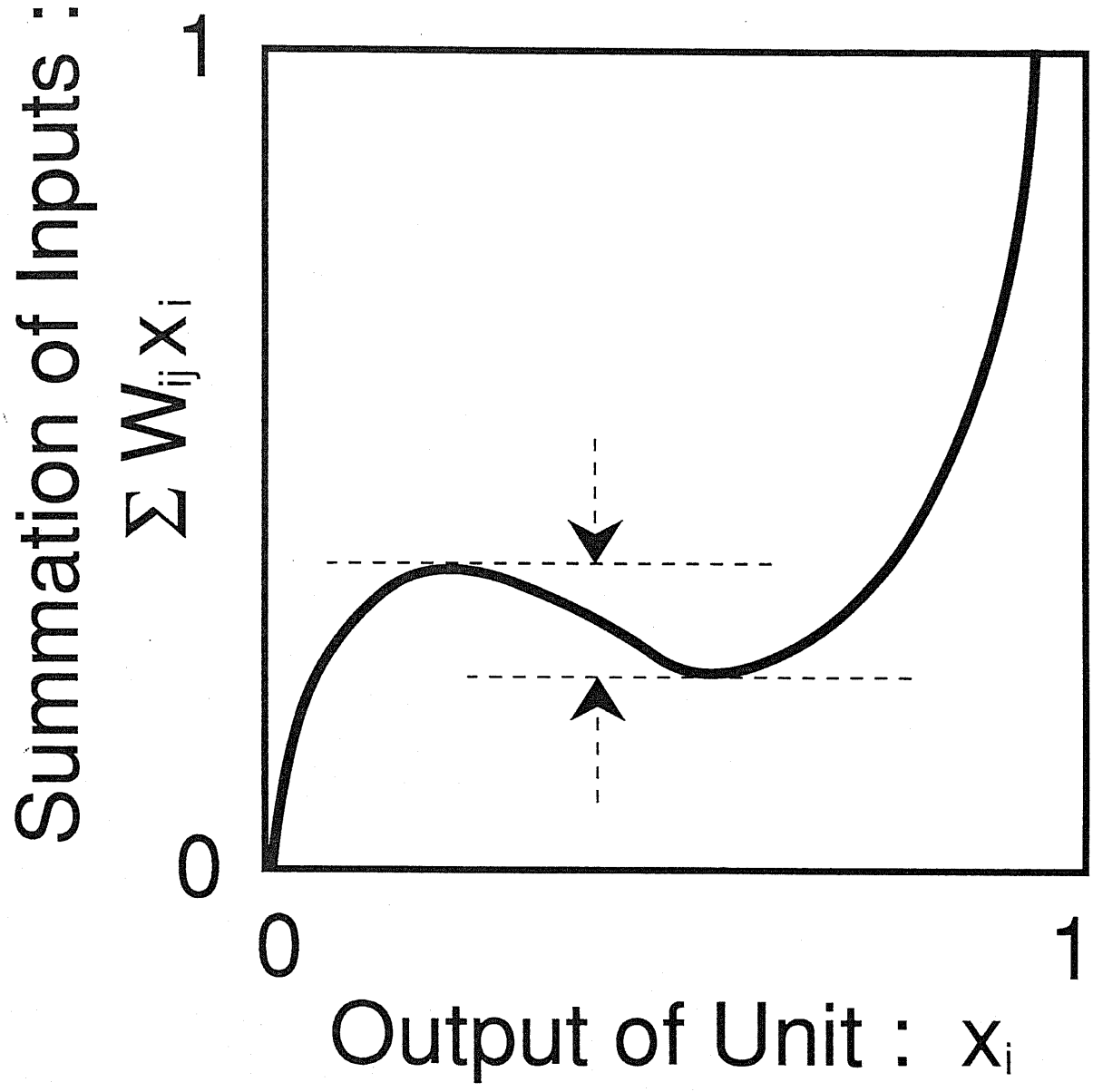


Fig. 3-12 Desirable Shape of Output of Unit

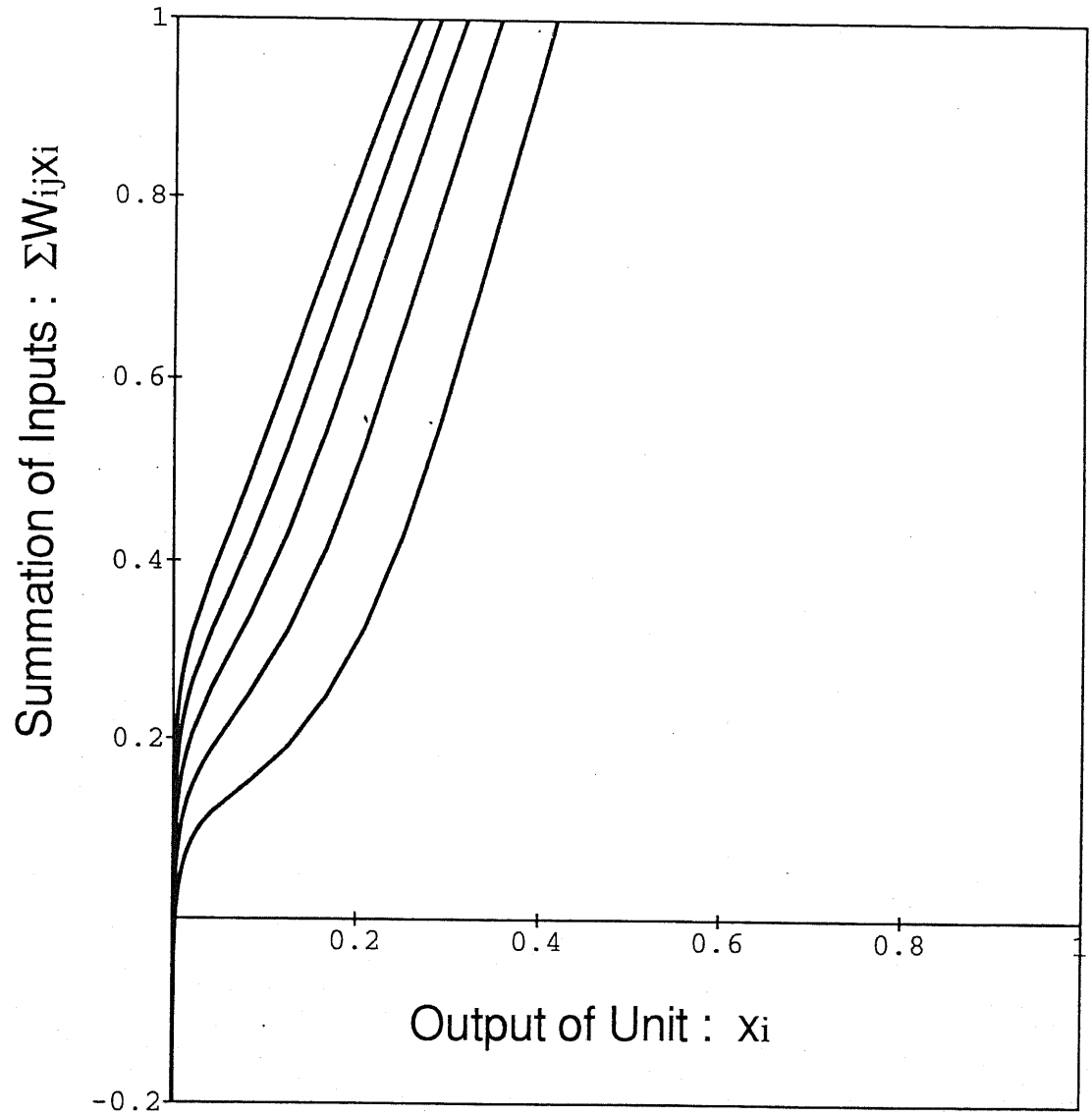


Fig. 3-13 (a) シミュレーションにおける特性を表す平衡曲線

$\alpha = 1.0$, $a_1 = 20.0$, $\theta_1 = 0.24$, $a_2 = 50.0$, $\theta_2 = 0.2$, $x_p = 0.25$, $y_p = 0.5$, NUNIT = 5

グラフ中上から順に y = 0.25, 0.20, 0.15, 0.10, 0.05

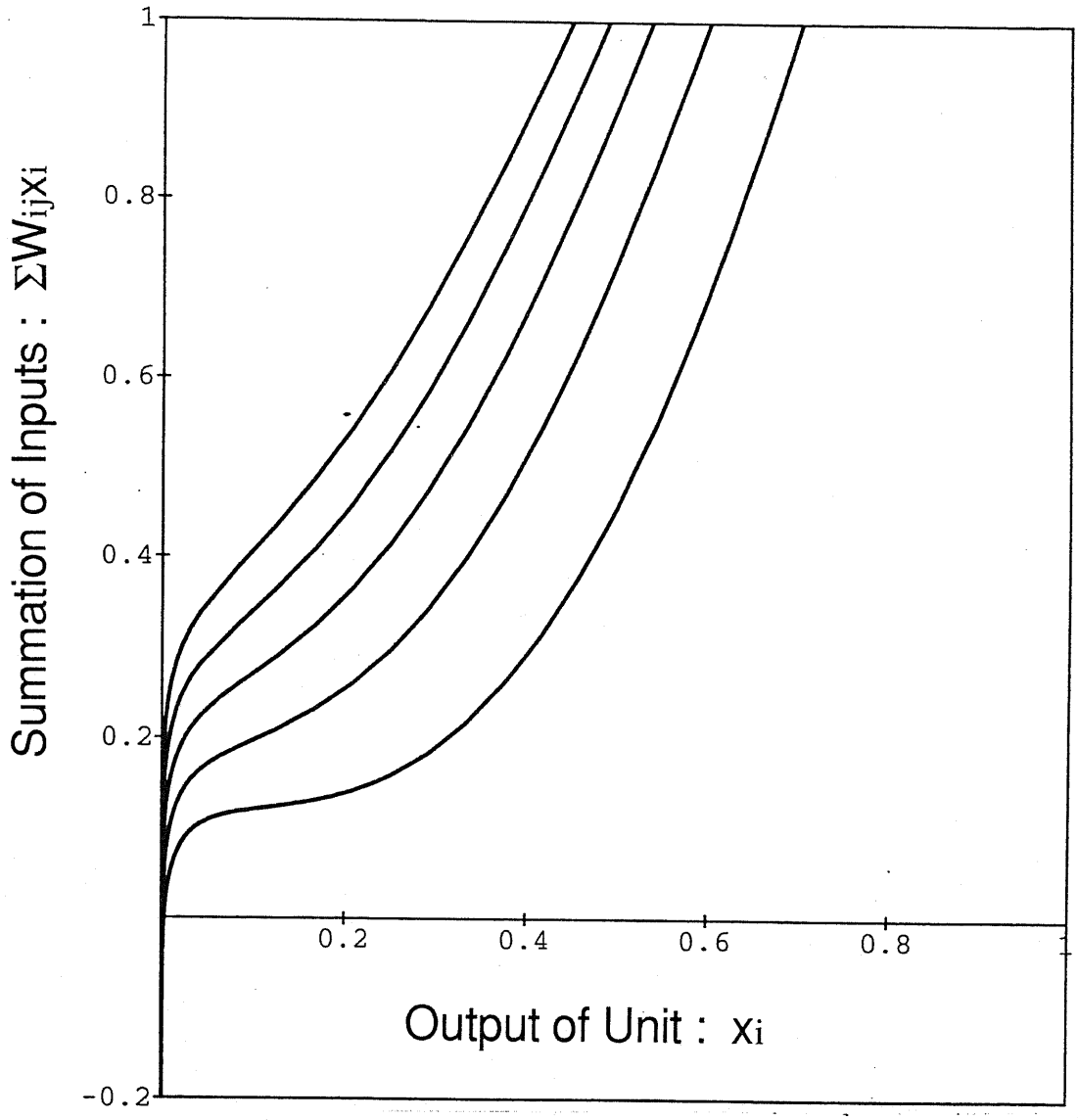


Fig. 3-13 (b) シミュレーションにおける特性を表す平衡曲線

$\alpha = 1.0$, $a_1 = 20.0$, $\theta_1 = 0.24$, $a_2 = 50.0$, $\theta_2 = 0.2$, $x_p = 0.25$, $y_p = 0.5$, NUNIT = 10

グラフ中上から順に y = 0.25, 0.20, 0.15, 0.10, 0.05

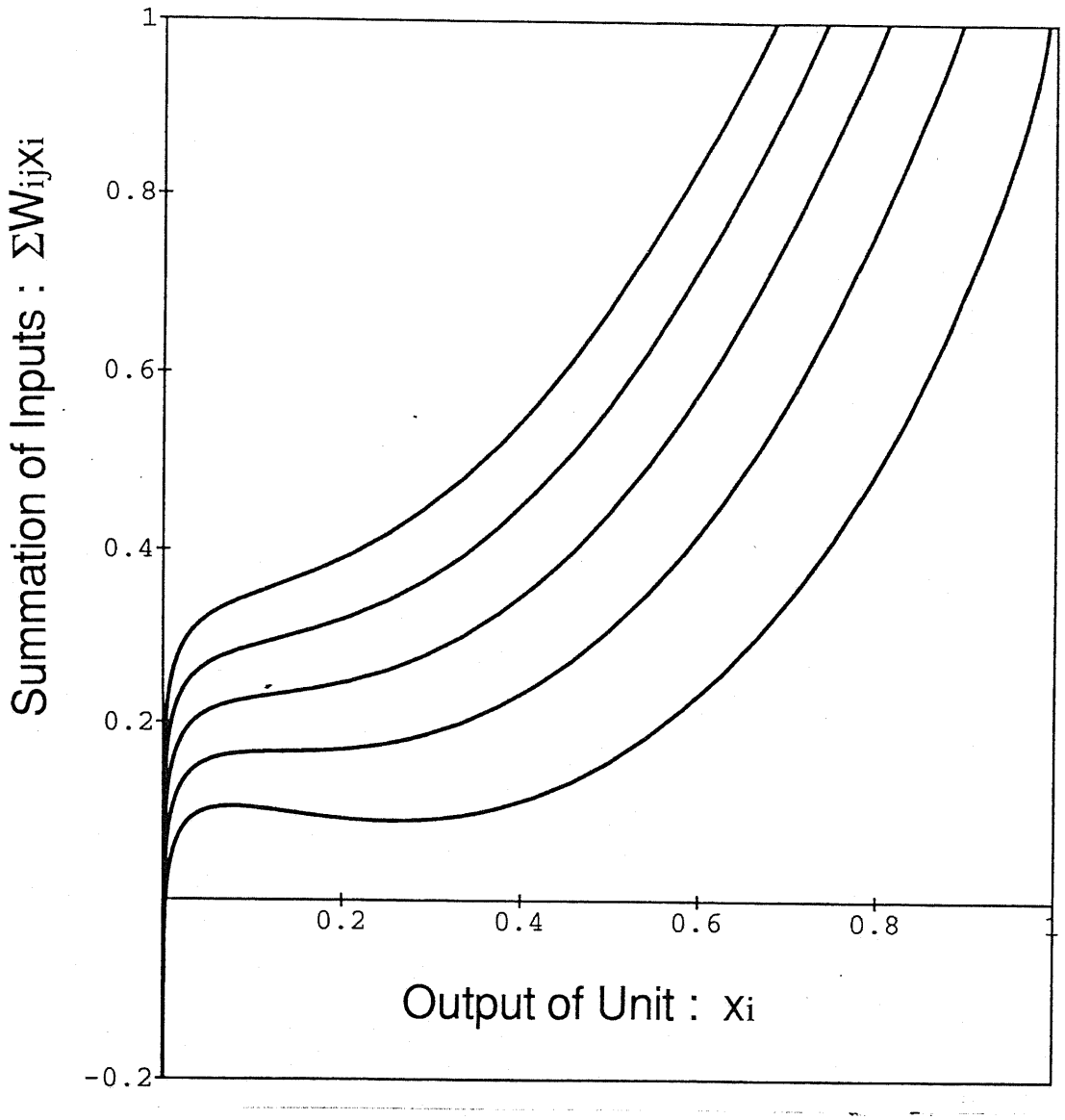


Fig. 3-13 (c) シミュレーションにおける特性を表す平衡曲線

$\alpha = 1.0$, $a_1 = 20.0$, $\theta_1 = 0.24$, $a_2 = 50.0$, $\theta_2 = 0.2$, $x_p = 0.25$, $y_p = 0.5$, NUNIT = 20

グラフ中上から順に y = 0.25, 0.20, 0.15, 0.10, 0.05

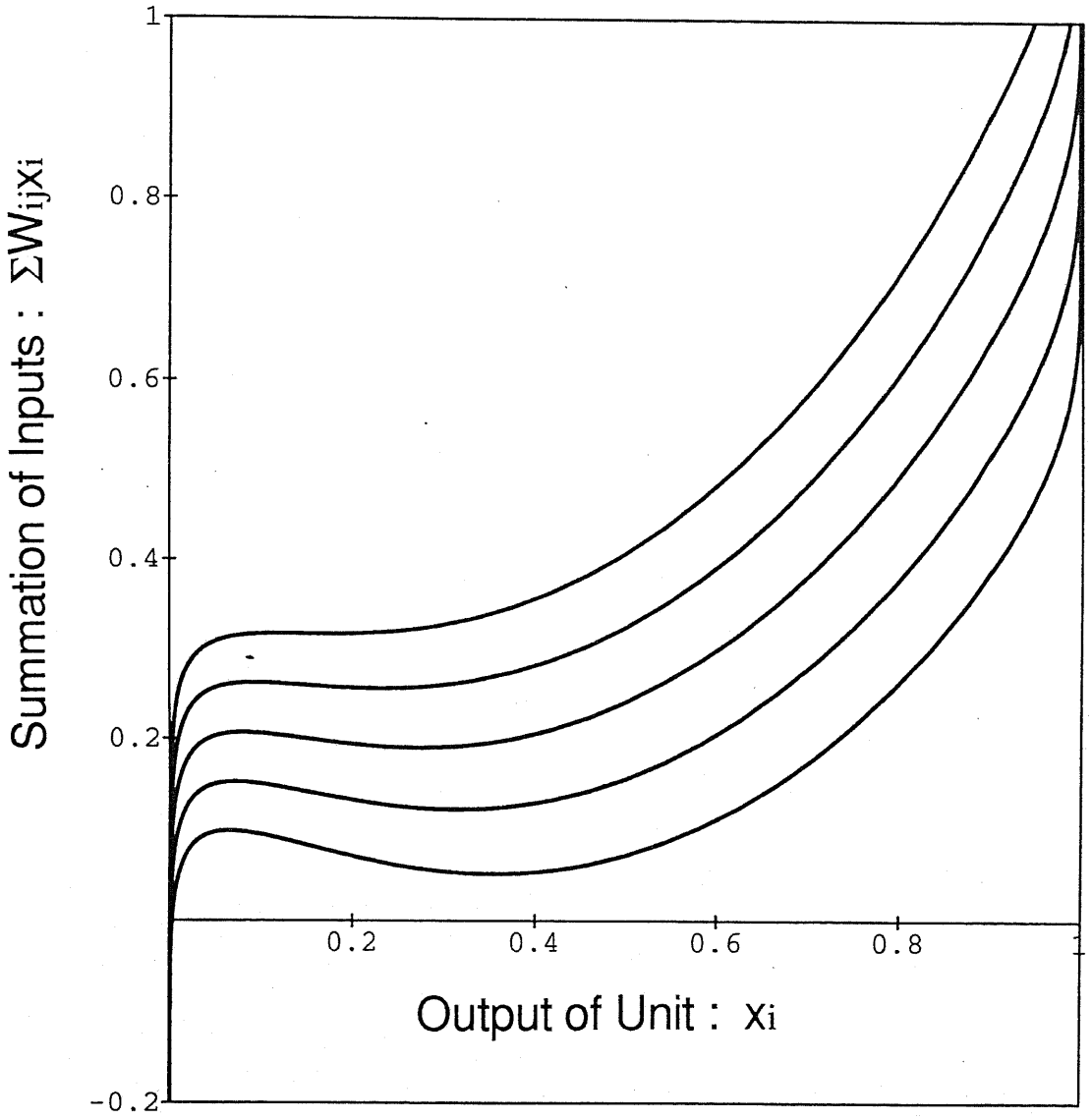


Fig. 3-13 (d) シミュレーションにおける特性を表す平衡曲線

$\alpha = 1.0$, $a_1 = 20.0$, $\theta_1 = 0.24$, $a_2 = 50.0$, $\theta_2 = 0.2$, $x_p = 0.25$, $y_p = 0.5$, NUNIT = 50

グラフ中上から順に y = 0.25, 0.20, 0.15, 0.10, 0.05

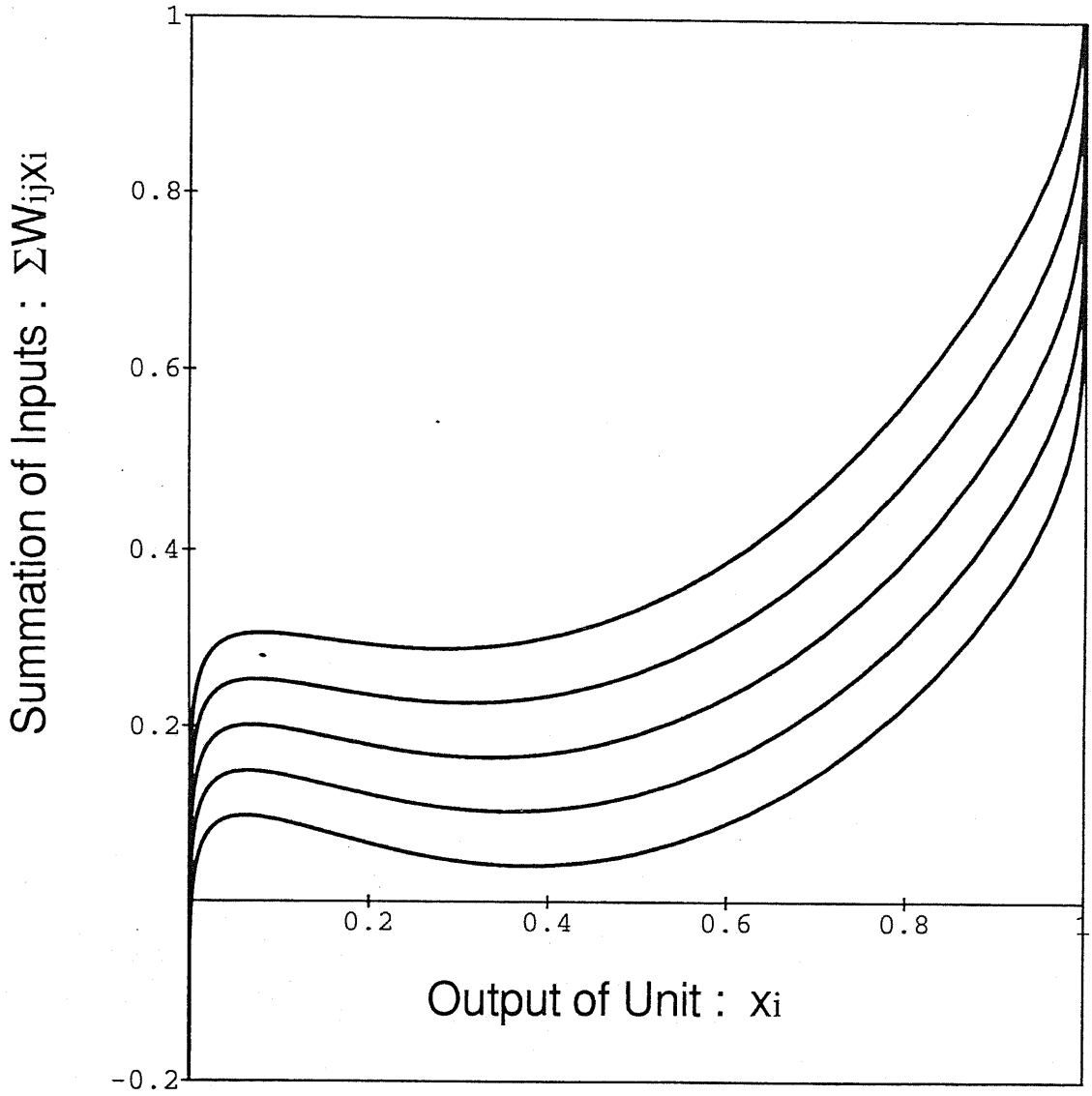


Fig. 3-13 (e) シミュレーションにおける特性を表す平衡曲線

$\alpha = 1.0$, $a_1 = 20.0$, $\theta_1 = 0.24$, $a_2 = 50.0$, $\theta_2 = 0.2$, $x_p = 0.25$, $y_p = 0.5$, NUNIT = 100

グラフ中上から順に y = 0.25, 0.20, 0.15, 0.10, 0.05

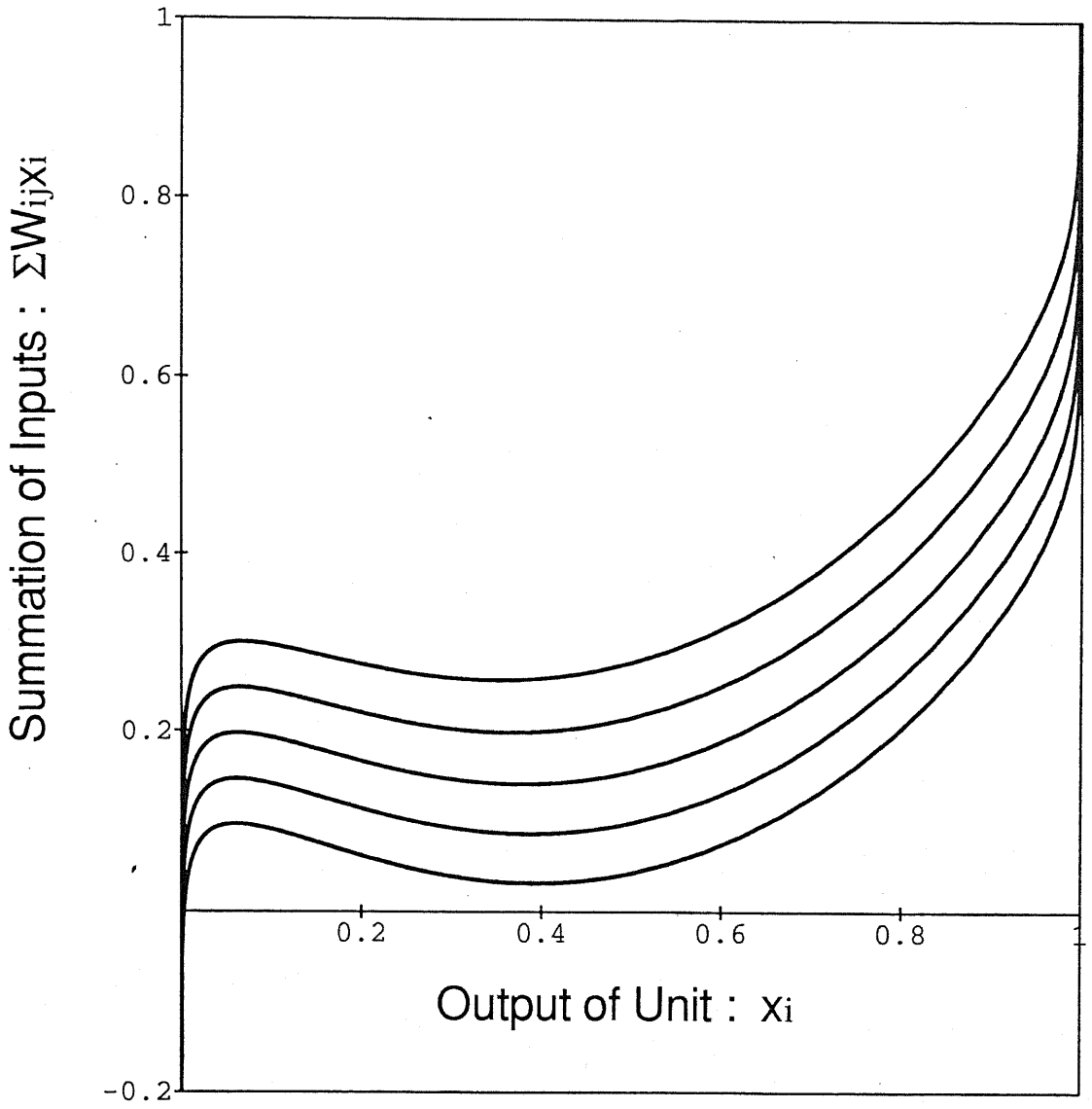


Fig. 3-13 (f) シミュレーションにおける特性を表す平衡曲線

$\alpha = 1.0$, $a_1 = 20.0$, $\theta_1 = 0.24$, $a_2 = 50.0$, $\theta_2 = 0.2$, $x_p = 0.25$, $y_p = 0.5$, NUNIT = 1000

グラフ中上から順に y = 0.25, 0.20, 0.15, 0.10, 0.05

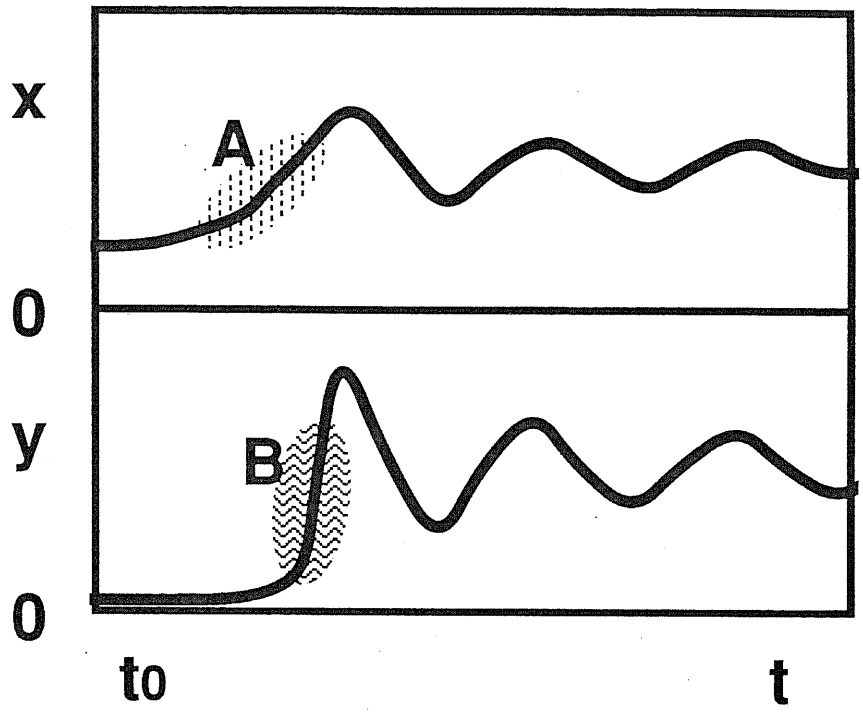


Fig.3-4 Problem of Fast Inhibition

第4章

迷路の中のニューロシステム

第2章で説明したニューロシステムを実現するためには3つのモジュールを開発する必要がある。そのうち出力ニューラルネットワークについては強化学習を利用したネットワークを利用すれば良いと思われるが、他の二つの価値評価モジュールと認識連合モジュールについてはより細かい研究を必要とする。このどちらを先に研究すべきであるかを考える。もちろん部分的にはどちらのモジュールの開発も可能ではあるが、全体のシステムとしてその効果をディスプレイするためには、Fig.4-1に示すような価値評価モジュールを取り扱ったシステムを研究すべきである。なぜなら、価値評価モジュールと出力ニューラルネットワークを備えたシステムはこれから説明するような迷路のような環境の中でゴールに到達する作業を効率良く行うことができるが。認識連合モジュールと出力ニューラルネットワークを組み合わせただけのシステムでは環境中における問題解決能力の向上を示すことは難しい。なお、価値評価モジュールの取り扱う価値観に対して、認識連合モジュールは世界観を取り扱うのだが、この世界観は推論や一般化を行うために役立つ。よって、私の個人的な考えでは価値評価モジュール自体には一般化能力を持たせる必要が無いと考えている。そのためシミュレーションで扱う環境として一般化を行うことが原理的に不可能な環境を用いた。多くの他の研究者による成果ではこの二つの機能を混在して取り扱かっているが、私はこの機能の明確な分離は少なくとも研究にとって、そしておそらくは生物のシステムにおいても重要な意味を持つと考えている。

以下、初期の研究とその問題点を指摘し、その後迷路のようなシミュレーションを行う。そこで価値評価モジュールの進化に伴って動作主体となるニューロシステムの能力が向上することを示し、さらには嫌悪性の価値観や確率的な環境においても、このシステムが有効であることを示した。

4-1. 初期の試みとその問題点

4-1-1. 報酬性価値評価モジュールのユニットコーディング

報酬性価値評価モジュールの研究の手始めとして、一般的なパターン間の相関ではなく、簡単なユニット間の相関に注目して研究をおこなった。シミュレーションに用いたモデルはFig. 4-2のように１又は0.を出力する10個の入力ユニットを持ち、一番目のの入力(x_0)に対して基本報酬発生器が出力(r)を行う。環境の設定としては、あるユニットに対する入力が一番目のユニットに対する入力と同時刻相関を持つとした。シミュレーションでは出力x_1およびx_3がx_0と同時に発火し、逆にx_2は同時に発火することは絶対にない。そして、それ以外のユニットはx_0に無関係なランダムに出力を行うものとした。よって、これを表現する式は

$$\Delta V_i = \beta \frac{d}{dt}\left(r + p - V_i x_i\right) x_i$$

$$\Delta V_i = -\gamma V_i \sum_{j=0}^{N} V_j$$

$$r = x_0, \quad p = \sum_{i=0}^{N} V_i x_i \qquad \text{(4-1), (4-2), (4-3)}$$

x_i : 入力ユニットの出力

V_i : 入力ユニットから二次価値発発生器への結合系数

t : 時間

r : 基本価値信号

p : 二次価値信号

N : 入力ユニットの数

β, γ : 定数

の様になる。式(4-1)は相関の学習で、式(4-2)式はそれに引き続く規格化である。また、あるユニットからの反応性が大きくなると自分自身の出力を帰還させてその結合が発散してしまう問題を避けるために、(4-1) 式では自己出力との相関を除くために$V_i x_i$が引かれている。

ここで、二次価値発生器に対する強化信号として$(r + p - V_i x_i)$を微分した量を用いたのは、もし$(r + p - V_i x_i)$自体を強化信号として使用した場合にはΔV_iが常に正となるので結合強められ続け、結果的に基本価値信号(r)と何等相関を持たないランダムな入力に対しても結合を強めてしまうためである。

シミュレーションの結果はFig.4-3の様で、ユニット単位のコーディングで同時刻相関に注目したモデルではその相関は学習された。

そこで、次の段階としてx0が発火する一時刻前に発火するユニットに対する結合を強めることを試みた。つまりユニットコーディングにおける一時刻ずれた相関が学習である。すると(4-1)式は時間を考慮して

$$\Delta V_i(t) = \beta \frac{d}{dt} \left(r(t) + p(t) - V_i(t) x_i(t) \right) x_i(t-1) \tag{4-4}$$

の様に書き換えられる。

ここで、前記のシミュレーション同様に微分を用いたが、ここに問題が発生した。つまり、現在価値があるとして反応しているパターンが減少したときにそのパターン自身の価値を弱めてしまう等の複雑な作用が発生するので単なる微分では相関をうまく学習することができないのである。

この問題を解決するために次のシミュレーションでは基本価値モジュールに対する入力結合の規格化の方法に工夫をして、静的な方法でもランダムな入力に対して結合を強めることがないように色々と調整を行い次の様な式に基づいてシミュレーションを行った。

$$\overline{x_i} = (1 - \delta)\overline{x_i} + \delta x_i$$
$$\Delta V_i = \beta (r + p - V_i x_i)\, \overline{x_i}$$
$$\Delta V_i = -\frac{1}{N} \left(\sum_{j=0}^{N} V_j - 1 \right) \tag{4-5), (4-6), (4-7}$$

$\overline{x_i}$: 時間に対して平均化された入力信号

δ : 減衰定数

その結果、一時刻ずれたユニット間の出力の相関をある程度うまく学習することができた。

結局、ユニットコーディングの範囲では次第に過去に遡って相関のあるユニットの活動に反応する再帰的な学習をすることが可能であることが示された。

4-1-2. 報酬性価値評価モジュールのパターンコーディングと問題点

前記の成果をより実際的な状況に適用するために、Fig.4-4に示すような2次元空間内の目標物を捕らえる課題に対するシミュレーションを行った。しかし、この場合先の技術ではうまく動作させることができなかった。この原因は前節でのシミュレーションがユニットコーディングであったのに対し、ここでの課題はパターンコーディングを必要としたからである。そのため、再び前節で行ったシミュレーションに対してパターンコーディングの場合にもきちんと動作するように研究を重ねた。

しかしここで、パターンコーディングの場合には、ある入力状態に対する価値評価がパ

ターン間の類似性に基づいて線型的に価値を予測してしまう問題点が明らかになった（本当はこれがニューラルネットワークの長所なのだが、ここでは障害になっている）。この問題点の端的な例について述べる。

例えばある状態Aから状態Bに移動するとゴールに近付き、状態Aとよく似たA'から状態Bに移動しても報酬性のゴールに近付く次のような場合

A -> B,　　A' -> B

状態に与えられる価値はゴールに近付くにしたがって大きくならなければならないので、A,A'に比べてBのに与えられる価値が大きくなるべきである。ところがAとA'が似通っているために、それらに与えられる価値が強化されてBよりも価値を大きく見積もられる可能性がある。この様な状態に陥ると動作主体はゴールに到達するための適切な動作を学習することができなくなる。

さらに、与えられる価値の大きさが入力ユニットの発火頻度のみによって与えられてしまう問題も発生した。

4-1-3. 解決のための方法

この問題を解決するための具体的な方法は二次価値発生器のに前段となる中間ユニットを導入し、それらユニットはそれぞれ単一のパターンに選択的に反応する様に学習を行うことである。この様なネットワークはニューロユニットの特性上実現可能である。つまり、二次価値発生器への入力の対象となる空間はユニットの数と同じ次元を持つ超空間であるが、この空間中の極めて局在した二つ以上の領域に対して反応することはニューロユニットにとっては不可能であるが、その領域が一つであれば容易に実現できる。そこで二次価値発生器の構成をFig.4-5に示すような二層構造とした。

以上のように、具体的には反応パターンの選択性の強化による方法で問題を解決できるが、逆にこの方法では類似した入力に対して同様の価値を与えるような能力は全く期待できない。つまり、一般化能力が無いのである。しかし、これはむしろ現在の研究段階においては有効である。つまり、これまでのこの種の研究では価値形成と一般化の課題の切り分けが不明快であったのに対し、この方法では一般化の問題を完全に無視している。言い替えれば、価値体系を作るという価値評価モジュールの働きは一般化の問題とは全く独立に取り扱える可能性があり、これら一般化や推論などの課題は我々の研究では認識連合モジュールの機能として今後の研究課題として残しておく。これらは、より一般的にはパターン推論の研究領域であろう。

そこで本研究では一般化や推論を行うことが不可能な環境を対象として動作主体の挙動

を調べることとした。その環境とは僅かの入力や出力の違いでも全く異なった意味を持つ非連続的（カオス的）な入出力関係を持つ空間である。

4-1-3. 迷路の世界への導入

前節で説明したように、ここでは一般化の問題を避けるために、一般化や推論を行うことができない環境を設定した。この目的に適した環境は、僅かな状態の差でも全く性質の異なる不連続な空間で、一般的な法則や傾向がないことが要求される。また、空間を連続的な時間や入出力パターンで記述すると無限個の状態を取り扱う必要があるので、これを避けるために時間と入出力パターンを離散的とした。この様に定義された空間は、おそらく動作主体に対する可付番な論理的空間である入力状態と、その状態と動作主体の行動出力の組み合わせ毎に定義される状態遷移ベクトルにより記述される。この様に定義される環境はFig.4-6に示すように、例えれば絵が描いてある各マス目と行動の種類によって分岐するマス目同志を連結する矢印によって構成される迷路の様なものである。ここで、動作主体に対する入力パターンはマス目毎の絵に、出力パターンはマス目間の移動動作に対応する。この環境が一般的な迷路と若干異なる性質は、第１に出発点が一定でないこと、第２に多くの場合移動動作は一方通行で後戻りができないことである（不化逆的：予測可能な等距離の逆行動が存在しない）。この空間は、見方によっては現実的な入出力空間から既に可能なあらゆる一般化を行ったサブ空間であると考える事もできる。

環境はさらに第１章で述べた、「遅れのある批判信号」を含む必要がある。このことは、環境中の特定の入力状態に反応して動作主体が報酬性またわ嫌悪性の批判信号をその内部で発生すると考えれば良い。これは、生物であれば摂食や危険回避に関する基本的な価値判断に対応する。

ちなみにこの視点で生物の行動を見てみよう。生物の目的とする最終ゴールは子孫の数を最大化するという取り留めのないものであるが、個体のレベルでは先天的な本能により基本的な部分については先天的に与えられている。つまり基本的欲求に対応する確からしいサブゴールが存在する。この様子はゴール周辺のみに方向を示す標識がある迷路ををさまよい歩いてゴールにたどり着くようなものである。

以下の議論では、環境を迷路のようなモノだと考えて議論を進める。

4-1-4. ゴール（報酬性）にたどり着く方法

動作主体が環境に対して適応するとは、できるだけ効率的に報酬性のゴールに到達することである。もちろん動作主体が環境に関する利用可能な知識を持ち合わせていなければ、可能な方法はとにかく適当に行動してみる（ランダム.サーチ）しかない。これを集団的に取り扱かい、総合的に成績の良いものを残す手法はジェネティック.アルゴリズムとして知られている。しかし、個体レベルでは動作主体はその行動により次第に経験を貯えるチャンスを持っているので、それらを利用すればより効率的な動作を行うことができる。究極的には、過去の経験を完全に貯えかつ利用する方法が思い当たる。つまり、動作の決定毎に全ての過去の経験を思いだし、しかもあらゆる種類の推論を行うのである。しかしこの頭でっかちな手法では経験を積み重ねるにしたがい情報処理の量が発散する。おそらく、この手法を用いたシステムは次第に考えているだけで何も行動しなくなるので、全く実用的でない。そこで、過去の経験を効率的で簡単な利用しやすい形式に蓄積する方法が求められる。

4-1-5. 距離標識と方向標識

例えば、始めて訪ねる目的地を目指してをドライブしているとき、どのような情報があれば能率よく目的地に到達できるのだろうか。通常役立つ情報とは、目的地までの距離を示した距離標識とその方向を示した方向標識である。この二種類の標識は、これまで議論してきた再帰的価値を自己形成するシステムと非常に対応がとれている。つまり、価値評価モジュールにおける価値観の形成は距離標識を立てることに、出力ニューラルネットワークにおける入力と出力の対応付けは方向標識を立てることに対応する。(Fig.4-7参照)ただしこの場合、価値の大きさはゴールからの距離に反比例する。

この２種類の標識の効果を明確にするために、その機能を説明すると、

方向標識　　[どの方向に進めばよいかを表示する]

次に行うべき行動を知ることができるが、今行った行動が適切であったかどうか判定できない。

距離標識　　[ゴールまでの距離を表示する]

次に行うべき行動が分からないが、行動前後の距離を比較することで、今行った行動が適切であったかどうか判定できる。

である。

そこで、これらの標識の一方だけを使用したシステムと両方使ったシステムの能力を考察し、併用した場合がいかに効果的であるかを示す。ただし、システムが短期的に記憶できるのは1時刻前の状態と距離のみであるとする。

(1). 方向標識だけ使用する

方向標識だけを利用した場合、ゴールに到達する動作を行ったとき以外は行動の良否が判定できない。同時に、行動の結果が適切かどうかを早い時点で判定できないと、何れの動作が効果的であったか推定することは難しい。よってゴールの近傍以外では正しい方向標識を立てられない。つまり方向標識だけでは行動の役には立つが、新たに標識を立てることが非常に難しい。このことは、新たに知識を貯えることがほとんどできないことを示している。

また既に多くのの方向標識が立っており、そのほとんどが正しいのだがいくつか間違いがある場合、どの標識が間違いであるのかを決定することができない。さらに、一度道に迷った末に方向標識がある場所に到達しても、ゴールに近付いたかどうかが分からない。この様に、方向標識だけでは十分な能力を発揮できない。

(2). 距離標識だけを使用する

一方距離標識だけを利用した場合には環境によりその有効性が異なる、環境が現実空間の道のように可逆的な行動を許す場合には、ある行動が距離を大きくした場合にその逆行動により元の位置に戻ることができるので、失敗を繰り返しながらも次第にゴールに近付くことができる。しかし環境が不可逆的である場合には一度大きくなってしまった距離は取り返しがつかないのでゴールに近づくことは事実上不可能である。

しかし、行動の結果により偶然にゴールや他の距離標識を発見すると一つ前のポイントの距離が推定できるのでそこに距離標識を立てることができる。

結局、距離標識は行動にはあまり役立たないが、新たな標識を立てることはできる。

(3). 方向標識と距離標識の両方を使用する

方向標識と距離標識を利用して効果的な行動と標識の設置を行う手だてを探ろう。

環境が可逆的ならば、ゴールから歩きつつ動作毎にゴールに向かう方向標識と1づつ大きい距離表示を立てる作業を色々な道筋について行うのが有効である。

不可逆で記憶が一つ前のポイントだけの場合に距離標識を立てるには、

『距離標識の無いポイントからのある行動の結果として距離標識を発見するか、または距離標識のあるポイントからの行動により一つ以上距離が減ったならば、直前のポイントに今行った行動を示す方向標識と距離を1大きくした距離標識を立てる。』 (Fig.4-7 参照)

という方法をとればよい。すると行動が成功するごとに次第に正しい標識を立ててることができる。

以上の手順で学習を行えば成功を重ねる過程で知識の範囲を拡大することができる。よって、非常に強力なシステムとなる。なお、もっと古いポイントまで覚えていることができれば、少し古い状態まで標識を立ててもよい。

この考察により、２種類の標識を併用する方法がいかに有効であるかが示された。この考察は第２章で導入された、再帰的に自己形成する価値体系の有効性に対する説明にもなっていることを、Fig.4-8を用いて説明しよう。ここでは環境と動作主体が相互作用を行っている。環境として第２章での例と同様にリンゴや栄養物の入力パターン（刺激）とこれらを結び付ける出力パターン（動作）に依存した遷移ベクトルが示されている。動作主体はFig.4-1の場合と同様に出力ニューラルネットワークと価値評価モジュールによって構成されている。出力ニューラルネットワークは方向標識に対応し入力刺激に対して適切な動作出力を行うように学習が進められる。価値評価モジュールは距離標識に対応し入力刺激に対して評価距離を出力する。学習前このモジュールは栄養物に対する距離が０であることのみが設定されている。

4-2. シミュレーションの説明

以上の様な仮定に基づくシミュレーションを行うために、より具体的な設定について説明する。

4-2-1. 環境の動作

迷路と考えた環境を設定することは入力パターン全空間中の遷移関係を記述する遷移ベクトルを行動パターン毎に決定すること、および動作主体にとってのゴールをある入力状態と決めることである。なお環境と動作主体間の信号は時間的に離散的な[0, 1]の二値出力であるとした。

遷移行列は環境がM入力N出力の系であるなら、一時刻前の出力状態の数：2^N個、動作主体からの入力状態の数：2^M個、現時刻の出力状態の数：2^N個を関係付ける$2^{(2N+M)}$個の要素を持つ遷移確率行列EP(x)により定義される。

$$\mathrm{Prob}(\mathbf{x}(t+1)) = \mathrm{EP}(\mathbf{x}(t), \mathbf{y}(t)) \tag{4-8}$$

x(t)： 環境の出力ベクトル ($x_1(t), x_2(t), x_3(t), \ldots, x_N(t)$)

y(t)： 環境の入力ベクトル ($y_1(t), y_2(t), y_3(t), \ldots, y_M(t)$)

t： 時刻

Prob()： ある状態が選ばれる確率

初期のシミュレーションで扱う環境は動作が確定的であると仮定するので、ある入力状態に対する出力状態は一意に決定する。よって、一つの出力状態の確率が1で他の場合は0となる。すると、上記の関係は次のように簡単化された式

$$\mathbf{x}(t+1) = \mathrm{E}(\mathbf{x}(t), \mathbf{y}(t)) \tag{4-9}$$

により表現される。ここで遷移行列(E)は、$2^{(N+M)}$の行列である。

さらに、2^N個の出力状態の中から、報酬性または嫌悪性のゴールもしくはその両方を一つ設定する。

シミュレーションにおいてはFig. 4-9に示すように、環境は2入力5出力の系としたので、一時刻前の出力状態の数：2^5個、動作による入力状態の数：2^2個、現時刻の出力状態の数：2^5個を関係付ける$2^{(5+2+5)}$個の要素を持つ遷移確率テンソルEP(x)を設定する必要がある。報酬性ゴールに相当する状態は5つの出力ユニットが全て1となる状態とし、嫌悪性ゴールに対応する状態は5つの出力ユニットが全て0となる状態とした。遷移行列は基本的に乱数により決定したが、孤立した状態が存在しないように、各状態から一つ番号が増える状態

への遷移路は必ず確保されるようにした。

本シミュレーションにおける遷移行列をTable 4-1に、そのダイアグラムをFig.4-10に示す。ここでの状態名は二進数の方法に基づいているので、31番状態が報酬性ゴールに相当し、0番状態が嫌悪性ゴールに相当する。なお、31番目の状態からの遷移は全ての状態に等確率で遷移するように設定されている。環境が確定的な場合には、遷移は全くこのとおりに起こるが、確率的な場合にはここに示された遷移の確率が主であるが、これ以外の遷移の可能性もある。

また、一般的には遷移ベクトルが経時変化をともなったり、未知の変数や隠れた変数を持つより複雑な環境を設定することができるが、本論文では議論しない。

4-2-2. 動作主体の動作

動作主体の構成はこれまで述べてきた強化学習システムを簡略化したものを用いる。ここでの研究の目的は価値評価モジュールの基本的な能力を調べることであるから、前処理的な働きを受け持つ認識連合モジュール(CAM)は取り除き、入力から動作を決定する出力ニューラルネットワーク(ONN)は簡単でしかも素直な性質を持っていた方がよいので、入力状態と行動に相当する出力状態を結び付ける確率テーブルに置き換えた。

4-2-2-1 出力ニューラルネットワーク(ONN)

出力ニューラルネットワークはFig.2-9で示したように、環境とは入出力関係が反転しN入力M出力のモジュールであり、シミュレーションでは2^5個入力状態と2^2個の出力状態を結び付ける。このモジュールは入力状態に応じた出力状態の選択確率を記述したテーブル(W)によりコントロールする。よって、この関係は、

$$\mathrm{Prob}(\mathbf{y}(t)) = \mathrm{W}(\mathbf{x}(t)) \tag{4-10}$$

x(t) : 動作主体の入力ベクトル = 環境の出力ベクトル $(x_1(t), x_2(t), x_3(t), x_4(t), x_5(t))$

y(t) : 動作主体の出力ベクトル = 環境の入力ベクトル $(y_1(t), y_2(t))$

t: 時刻

Prob() : ある状態が選ばれる確率

と、表される。

初期状態においてこれらの確率は4種類の動作を等確率で選ぶように設定しておく。学習則は基本的に強化学習であるが、確率テーブルを使用する為にその規則は若干変更を加

えてある。当初のシミュレーションでは、環境が確定的であるからONNの学習規則も最も単純な方法をとった。つまり、SVAMから強化信号が到達したならば、一時刻前と同じ入力に対しては常に直前の動作を行うように学習を行う。つまり、学習が行われたのと同じ入力に対する動作の仕方は決定論的になり、一つの出力状態の確率が1となり、他の出力状態に対する確立は0になる。

4-2-2-2. 価値評価モジュール(VAM)

一方、価値判断モジュール(VAM)は、入力状態に価値を与え、その変化から強化シグナルを生成する機能を持つ。これは前記した距離標識に当たり、ここでの価値基準はゴールへの距離として表現される。このモジュールは二層の構造を持ち(Fig.4-9参照)、入力側の第一層は複数個のユニットにより構成され、第二層は単一のユニットである。それぞれの機能は

第１層の動作

1). 一つの特定の入力状態に選択的に反応する。
2). その入力状態のゴールへの距離を出力する。

第２層の動作

1). 第１層の出力を集めて距離を調べる。
2). 距離の変化から強化シグナルを生成する。

ここでは基本的なコンセプトにおける原始価値モジュール(PVAM)と二次価値モジュール(SVAM)は並列に配置されており、原始価値に相当する機能はゴールの状態に反応するようにあらかじめ変数が設定された一つの先見的なユニットで、それ以外のユニットが学習により次第に有効な数を増やす二次価値に相当する。

4-2-2-3. 学習規則

この後システムの学習規則について検討するが、しばらくのあいだの議論では報酬性のゴールへの評価距離の変化によってどの様に学習を行うかを検討し、嫌悪性のゴールについては後に譲る。

学習を行う部分は、価値評価モジュール第一層の入力選択性と、その出力の距離評価量および出力ニューラルネットワーク(ONN)の確率テーブルである。学習が行われる条件は現時刻(t)および一時刻前(t - 1)に価値評価モジュールによって生成された報酬性ゴールまでの評価距離($D_R(t)$, $D_R(t-1)$)によって規定される。

価値評価モジュールの学習規則の一つは、生成規則による。

生成規則

ある反応すべき入力パターンに反応するユニットがなかったら新たにその入力パターンに反応するユニットを生成し、適切な評価距離を設定する。

具体的には価値評価モジュールの第１層のユニットの一つを有効にし、直前の入力パターンに対する反応性を学習し、現在の評価距離よりも一つ大きい値をそのユニットの出力すべき評価距離として設定するものである。この学習が行われるのは、一時刻前には反応するユニットがなく距離が評価されず、現時刻において評価距離が判った場合である(Table 4-2中のCreateSvam)。

また、一時刻前の評価距離より現時刻の距離が１以上小さいときには前時刻に反応したユニットの距離評価を（現時刻における評価距離 ＋ 1）に変更する必要がある(Table中のSetDist)。ただし、環境と動作主体がともに確定的に動作しているこの段階のシミュレーションでは実際には効果がない。

出力ニューラルネットワーク(ONN)の学習は価値評価モジュール(VAM)の強化信号により制御される。強化信号が出力される必要があるのは、報酬性ゴールまでの距離が減少したときである(Table 4-2中のLearnAct)。これにより、ONNは直前に行なったの行動を強化することができる。この条件が揃うのは、Table 4-2に示すように、一時刻前には距離が判らなかったが現時刻において距離が評価できた場合である。一時刻前に距離が判っていて現時刻でさらに距離が縮んだ場合も考え得るが、環境と動作主体がともに確定的に動作しているこの段階では実質的にはやはり効果がない。

		$D_R(t-1)$ known		$D_R(t-1)$ unknown
		$D_R(t-1) \neq 0$	$D_R(t-1) = 0$	
$D_R(t)$ known	$D_R(t) < D_R(t-1)$	**SetDist** LearnAct	None	**CreateSvam** LearnAct
	$D_R(t) = D_R(t-1)$	Impossible		
	$D_R(t) > D_R(t-1)$			
$D_R(t)$ unknown				None

Table 4-2　完全に確定的な場合に学習を行う条件

確定的環境中における確定的な動作主体の報酬性ゴールに対する学習規則。

$D_R(t)$, $D_P(t)$:　現時刻の報酬性ゴールからの距離。

$D_R(t-1)$, $D_P(t-1)$:　一時刻前の報酬性ゴールからの距離。

CreateSvam : (Create SVAM) 価値評価モジュール内の一つのユニットを有効にし、そのユニットの保持する評価距離を、(現在の評価距離 + 1) に設定する。

SetDist : (Set Distance)　一時刻前に反応した価値評価モジュール内のユニットの保持している評価距離を、(現在の評価距離 + 1)に書き換える。

LearnAct : (Learn Action)　直前に行った行動を強化するように出力ニューラルネットワークの学習を行う。（強化信号が出力される）

None : (None)　何も学習しない。

Impossible : 発生し得ない条件。

以上の学習規則に基づいて学習が進行する様子をFig.4-11に基づいて説明する。左上には報酬性ゴール（状態31）周辺の状態遷移ダイアグラムが示してあり、例えば状態７から行動２を選択すればゴールに到達することができる。ここで、各状態の下に付属している５つのマス目は動作主体に対する入力刺激のパターンを表している。その右に行動を制御する出力ニューラルネットワーク内の確率テーブルと価値評価モジュールに貯えられた評価距離のテーブルの初期状態が示されている。ここに見られるように動作が四つであるからその選択確率は各々1/4で、評価距離についてはゴール自身の距離が0であること以外は判っていないので-1で表してある。その下の中段では動作主体が始めて状態７から行動２を選んだ結果ゴールに到達したときに出力ニューラルネットワークと価値評価モジュールのテーブルがどの様に書き換えられるかを示した。ここで、入力状態７に対しては行動２を選ぶことと、状態７のゴールへの距離が１であることが学習された。この様な事態は次に状態６から状態７へ移動した時にも起こることが下段に示されている。この連鎖は次々と起こるので結果として動作主体は環境に対する適切な行動と価値観を修得することができるのである。

4-2-3. 具体的な処理の流れ

本シミュレーションに用いたプログラムおける主な処理は次にあげるものである。

1. 環境の更新
2. 価値評価モジュールの更新
3. 出力ニューラルネットワークによる行動出力の決定
4. 価値評価モジュール第１層の生成、再学習
5. 出力ニューラルネットワークの学習

4-2-4. 評価の方法

4-2-4-1. ゴール到達確率

シミュレーションのある時点における動作主体の能力を見積もるためには、学習を止めたまま同じ環境中で動作させた場合に、どの程度ゴールに到達できるかを調べれば良い。この際シミュレーションと同様に動作させる方法もあるが、むしろ動作主体がある状態に

存在する確率を考えた方が簡単かつ包括的な評価が行なえる。

以下に私が採用した方法を説明する。まず、任意の時刻(t)における存在確率ベクトル

$$P(t) \equiv (p_0(t), p_1(t), p_2(t), \ldots\ldots, p_{31}(t))^T \tag{4-10}$$

により、各状態に存在する確率が表されたとすると、これは環境と動作主体内部の変数により決定される遷移行列をTに基づいて変化し、その一時刻の状態変化は

$$P(t+1) = T\,P(t) \tag{4-11}$$

$T = \{tij\}$: 1step での状態変化をあらわす遷移確率行列

i: final state(i = 0~31), j: initial state(k = 0~31)

$$t_{ij} = \begin{cases} = \sum_{k=0}^{31} a_k EP_{ikj} & (\text{if } j \neq 31) \\ = 0 & (\text{if } j = 31) \end{cases} \tag{4-12}$$

ただし

ai : 動作主体の動作選択確率(i = 0~3)

EPikj : 環境をあらわす遷移確率行列,

i: final state(i = 0~31), k: action(k = 0~3), j: initial state(j = 0~31)

と記述される。

さて、動作主体が時刻0において状態0にあるときその分布は

$$P(0) = P_0 \equiv (1, 0, 0, \ldots\ldots, 0)^T \tag{4-13}$$

と、表される。この状態から時刻tにおける存在確率分布を求めると

$$P(t) = (T)^t P_0 \tag{4-14}$$

である。よって、その時にゴールに到達する確率は$p_{31}(t)$である。動作主体の性能の評価基準としては、この到達確率をゴール以外の全ての出発点について平均すべきである。この値は

$$\overline{P(t)} = \frac{1}{31}\sum_{i=0}^{30} (T)^t P_i \tag{4-15}$$

ここで、一次変換は線型変換であるから

$$\sum_{i=0}^{30} (T)^t P_i = (T)^t \sum_{i=0}^{30} P_i \tag{4-16}$$

つまり、

$$\overline{P(t)} = \frac{1}{31}(T)^t \sum_{i=0}^{30} P_i \tag{4-17}$$

となるので、計算においては初期の状態においてゴール以外の状態に等確率分布している状態を設定し、その後の先ほどと同じように一時刻毎のゴールに到達する確率（$p_{31}(t)$）

を計算すればよい。

以上の手順によりステップ毎のゴール到達確率が算出できる

4-2-4-2. 平均ゴール到達ステップ

上記の基準有用であるが、いくつかのモデルの能力を比較する場合などには一次元の量が好ましい、そこで距離をゴール到達確率で積分することにより平均何ステップでゴールに到達できるかを示す平均ステップ数を

$$\text{Average Steps} = \sum_{t=1}^{\infty} \overline{p_{31}(t)}\, t \tag{4-18}$$

のように算出することができる。

現実の計算においては無限のステップ数を計算することは出来ないので、ある距離まで積分して打ち切りを行う必要がある。ここで、確率をあらわす係数$p_{31}(t)$は（後に示す結果の図にも示されるように）距離に対して指数関数的に減少するので、距離と確率の積は遠方でのの寄与は極めて小さくなるので、この打ち切りは多くの場合問題がない。

4-3. 基本モデルから再帰的モデルへのシミュレーション

まず第２章で説明した、基本モデルから再帰モデルの開発に沿って同一の環境に対し、これらのモデルを適用した場合にその能力にどの様な差が現れるかを計算機シミュレーションによって検証した。環境は動作主体にとって適応することが容易な確定的及び不変な状態遷移関係をもつものとする。

4-3-1. 基本モデル（０次モデル）

2-3-1-1.で説明した、二次価値評価モジュールをもたない動作主体に対してシミュレーションを行った。

環境を一定に保った上で動作主体の確率的な行動選択に関わる乱数の種のみが異なる３通りの結果についてFig.4-12, Fig.4-13, Fig.4-14に示す。それぞれの上図は前節で説明したゴール到達確率が示してある。グラフ中のそれぞれの曲線は全く学習をしていない学習ステップ０から60ステップ環境中をさまよい歩く毎の能力を示してある。一方下図は動作主体が環境中をさまよい歩く過程で何回報酬性ゴールに到達したか、また前節で説明した報酬性ゴールへの平均ゴール到達ステップ数、さらに上図の確率のピーク位置そして報酬性価値評価モジュール内の中間ユニットの数を示している。

初期状態におけるゴール到達確率は各上図の中の一番下の曲線であり、１ステップでゴールに到達する確率は1/31(=0.032)である。（ゴール以外の状態の数が31個ありゴールに直結する経路が4つあり、動作主体がある行動を選ぶ確率が1/4であるので。）学習に伴って次第に少ないステップ数でゴールに到達する確率が増大する。しかし、このモデルでは動作主体は直接ゴールに到達する行動しか覚えることができないので到達確率が最も大きいのは学習後においても１回の動作ステップでゴールに到達する確率である。シミュレーションに用いた環境ではTable 4-1やFig.4-10に示すように、ゴール（状態31）に到る経路は状態6,7,19,20からの４通りであるから、このモデルが可能な範囲で完全に適応するには、4つの経路を覚える4段階のレベルがある。３つの図中の上図ではそれぞれ3,4本の曲線しか表示されていないが表示ステップのきざみをより細かくとれば、5本の曲線が表れるはずである。ただし、Fig.4-12の例ではまだ３つの経路しか学習していないので一本足りないはずである。報酬性ゴールに到達した回数は学習の進行に伴って平均到達ステップ数がが小さくなると急増する。平均到達ステップ数は最終的に約9になり、ゴールに到達する回数を調べてもやはり50回到達するのに400から500ステップを要するようである。なお報酬性価値評価モジュール内の中間ユニットの数は二次価値を扱わないので初期に与えた基

本価値評価に関するユニット１つが変わらず存在し続けるだけである。

平均到達ステップ数をの変化を見ると、ゴールに到達したからと言って必ずしもその能力が向上しないことが判る。これは、様々な経路からゴールに達っしたならば多くの経路を学習できるが、いつも同じ経路ばかり通過していると例えゴールに到達することができても新たな知識を修得できないためである。

4-3-2. １次モデル

2-3-1-2.で説明した一次モデルではゴールの一つ手前の状態に価値を与えることができるので、Fig.4-15, Fig.4-16, Fig.4-17に示すようにゴールに到達する能力が向上する。この場合も基本モデルと同じ環境に対する３通りの場合について示している。

学習は基本モデルよりも素早く進行し、しかもゴールに到達する能力は基本モデルよりも遥かに優れている。上図の初期状態は当然基本モデルの場合と同じである。学習後は2ステップでゴールに到達する確率が最大となるのは、ゴールの一つ手前の状態に到達する行動を学習することができるためである。ゴール到着確率のグラフが上に凸型になるのは１ステップでゴールに到達することができる入力状態よりも２ステップでゴールに到達できる。状態の数が多いためである。下図を見ると基本モデルでは50回程ゴールに到達するの600ステップを要したのに対して本モデルでは400ステップほどで達成されている。最終的な平均ゴール到達ステップ数は約4ステップとなる。しかし、これらのモデルでは何れもゴールに直接到達する経路は３つしか学習されていないことは、増加した中間ユニットの数が3 つであることからも明かである。

4-3-3. 再帰モデル

さて、第２章で開発された（2-3-3.参照）再帰モデルについてのシミュレーション結果を見てみよう。これまでと同じ環境の基で乱数の種だけを変化させた5通りの結果をFig.4-18~Fig.4-22に示す。ただし報酬性距離を覚えることができる距離の範囲に制限を加え、その上限を3とした。

各上図を見ると一次モデルよりもさらにゴールに到達する能力が向上していることが判る。つまり、より短いステップ数でゴールに到達する確率が増え、10以上のステップを要

する確率は急激に減少する。しかしゴールに直接到達する4つの経路を全て学習した動作主体はこの例の中には無い。到達確率がピークとなるステップ数は3又は4となる。興味深いのはFig.4-18, FIg.4-19, Fig.4-20では到達確率の形状が自然な上への凸上であるが、Fig.4-21, Fig.4-22では3ステップで到達する確率が2ステップや4スッテプで到達する確率よりも小さくなる凹型の形を含む点である。これは、後者の動作主体がステップ数の少ない簡単な経路よりもステップ数が4つほど必要なより複雑な経路を得意としてしまったことを示している。各下図に目を移してみると平均到達ステップ数は若干一次モデルよりも改善されている。中間ユニットの数は何れの例でも最終的に上限の10個となっている。

4-3-4. 考察

これらシミュレーションの結果、何れの場合においても、動作主体は一度ゴールに到達するまではランダムに動き回っているだけなので非常に苦労するが、ある程度環境の構造がわかりはじめると急激に動作能力が向上する。特に初期の数回のゴール到達による学習の成果が能力の向上に大きく寄与していることが、平均ゴール到達ステップの変化からよくわかる。

再帰モデルが一次モデルに対してあまり能力の向上を示すことができなかったのは、環境に原因がある。例えば全ての入力状態が直接ゴールに結合しているような環境の場合には基本モデルで十分であり、むしろ高次のモデルを用いると必要以上に遠回りをする経路を学習して適応度を落すことになる。このことは再帰モデルと一次モデルの場合にも当てはまる。つまり環境によって適切な学習距離が存在する。実際、これらモデルの上図の結果を良く比べてみると、１ステップで到達できる確率に関しては基本モデルが最も優れており、２ステップで到達できる確率に関しては一次モデルが最も優れている。そこで、この学習可能な距離の自動制御は今後に残された課題である。

学習された動作パターンは高次なモデルほど、独自な得意のゴール到達パターンをマスターし、それを多用する傾向が強くなる。つまり、経験に基づいた個性が表れる。これは、第１章に示した本研究の一つに目標に答える結果である。

つぎに、ゴール到達確率（それぞれの上図）には対数表示を用いてきた。何れのグラフを見てもあるステップ以上ではほぼ直線的に減少していることが判る。これはステップの増加に伴って到達確率が指数関数的（exp(-bn)）に減少していることを表している。この種の振舞いは一般に流量と残量が比例する容器からの流量の時間変化の場合に相当すると

考えることができる。ここでは、容器としては環境内の動作主体にとって未知の領域が相当し、そこから既知の領域へ状態が変化する量によってこの傾きが決定される。つまり

exp(-b) = 未知の世界から既知の世界への遷移確率

とである。

さらに状態間の遷移確率を平均化して考えれば、

$$\exp(-b) = Nul / Nl \tag{4-19}$$

Nul : 既知の状態の数

Nl : :未知の状態の数

そこで、動作主体が全環境に対してどの程度の割合について行動の知識を持っているかを見積もる量として環境に対する既知率(r)を定義すると

$$r = 1 - \exp(-b) \tag{4-20}$$

とすることができる。

各シミュレーションの最終状態でのこの値は、基本モデルでは0.1程度で、一次モデルでは0.3程度、再帰モデルでは0.4程度であるが、再帰モデルではかなりのばらつきがあり、再帰モデル#5では0.56に達する。ただし、既知率は行動方法を知っている状態数の割合であるが、そこで知られている行動は必ずしも最適な行動とは限らない。

4-4. 嫌悪性価値観

これまで、報酬性価値観および報酬性のゴールに到達する方法についてシミュレーションを行ってきたが、本節では嫌悪性の価値観を取り扱い嫌悪性ゴールを回避する方法についての研究を説明する。

4-4-1. 嫌悪性価値観の説明

生物における嫌悪性の最終的なゴールとして死を想定すると、死は動作主体自体が機能を失うことを意味するので、死んで始めて間違いに気づいたのではもはや何も学習することはできない。そこで、最終的な嫌悪性ゴールに達する前の警報や警告の段階を嫌悪性のゴールとし、この段階で回避行動を行うように適応を行っている。つまり報酬性との比較において、「快い事はとことん試す事ができるが、危ない事は最後まで試す事ができない」という差がある。

次に、ゴールからどの程度の距離までその距離を学習すべきかを検討する。報酬性ゴールに関してはゴールからの距離が大きくてもその距離を学習することは比較的有用であった。しかし、嫌悪性ゴールの場合、あまり遠くから距離を学習すると、その効果により多くの入力状態に対して回避行動が固定化され、動作主体が探索できる範囲を自ら狭め、結果として動作主体の能力を落す場合がある。しかしここにはトーレードオフの関係が存在する。もし環境が嫌悪性ゴール直前の状態に必ず回避経路を持ち、しかも確定的に動作するならば、動作主体は嫌悪性ゴールからの距離は１まで学習するだけでほとんど十分である。しかし、環境が確率的であったり、嫌悪性の袋小路を持つならば、学習可能な嫌悪性距離をより大きくとる必要がある。本節のシミュレーションでは嫌悪性袋小路のない確定的な環境を対象としているので、学習可能な距離は確定的な１としてあるが、今後検討を要する微妙な問題である。

つぎに、出力ニューラルネットワークが行う強化学習は報酬性の場合と同じで良いか検討する。すると、報酬性の価値観を学習する際には明かだった強化学習規則が嫌悪性に関しては必ずしも明確に定義できないことに気づく。この不明確さの原因は、報酬性では価値を形成したときの行動と学習すべき行動とが一致したのに対し、嫌悪性ではそうではないことに由来する。そこで、３種類の嫌悪性に関する学習則の候補を検討した。

その１.逆固定型（逆のパターンを強化する）

行動により嫌悪性価値が出現または増大した場合に、直前に行った出力パターンとパターン的に逆向きに強化を行う。

この方法は報酬性における強化学習をそのまま嫌悪性に延長した場合に相当するので、ニューラルネットワークで容易に実現可能である。しかしこの方式はFig.4-23に示すようなパターン的に逆の状態がともに嫌悪性ゴールに近付く動作となるような状況に出会ったときに、振動的にこのお互いに逆のパターンの行動を学習することになり、永久に行動が改善されないという問題が発生し得る。二つ目の問題として、Fig.4-24に示すようにある状態(A)において報酬性ゴールに近付くためのより優れた行動の選択支があるにも関わらず、それ以外のたいして意味のない行動を固定化してしまう難点がある。この問題点は時にある困難から逃れるために再び嫌悪性のゴールに近付くような悪循環を形成する可能性を高める。

その２.脱出記憶型（脱出に成功した行動を記憶する）

行動により嫌悪性が消滅又は小さくなったときに、直前の行動を強化する。つまり、危なくなった段階で色々な行動を試し、結果として危険な状態から脱する事ができたならば、その行動を強化する方法である。

この方法では嫌悪性を感じたときにランダム性を増大させて、これまでの習慣に引きずられないようにする事が必要である。さらに、この方法は他の二つの学習則とは、行動学習のタイミングが異なるため、他の学習則と同様のレベルの回避行動を行うためには一つ遠くの距離標識が必要とされる。構造的にはやはり報酬性の場合のニューラルネットワークのアーキテクチャーをそのまま使うことが出来るので、都合がよい。

この方式では前記方式における、一つ目問題点は解決できるが、二つ目とほぼ同様の問題点をもつ。Fig.4-24において入力状態のパターン的な意味は特に考えないこととしても、状態Ａから行動３により一度状態Ｃに到達すると状態Ａに嫌悪性価値観が付加されるので、その後状態Ａに到達したとき行動１を選べばその回避行動が有効であったと認められ行動１が強化される。つまりより優れた行動２を学習する可能性が奪われてしまう。もちろんこの振舞いはは一概に欠点だとは言えないが、動作主体にとってある意味で不利益な特性である。

その３.失敗否定型（失敗した行動だけを否定する）

行動により嫌悪性が出現又は増大したときに、直前に採った行動だけを否定し次には行わないようにする。つまり、同じ失敗だけは繰り返さないようにする方法で

ある。

この方法は、その機能においては最も望ましいが、ニューラルネットワークで実現するには例えば、直前の行動出力に対して選択的に反応するユニットを形成し、その出力によってその種の出力を抑制するなどの構成を必要とする。しかし、この方法では否定したいパターンの数だけニューロユニットを準備しなくてはならないので、生体システム内で実現することは不可能であると思われる。また、工学的な観点からも大量のメモリーを必要とするために好ましくない。

これら三つの方式を比べてみると逆固定型は一つ目の問題が致命的であるし失敗否定型はあまりにも多くの記憶容量を必要とするので、実際的には脱出記憶型が有効であると思われる。さらに脱出記憶式が問題が人間にもっとも近い戦略であると思われる。そこで、今後、この方法を中心に議論を進めてゆく。また本章の議論では一般化の問題を無視しているが、一般化を考慮した場合には逆固定型と脱出記憶型の混合した方式だと考えられる。

これまでの多くの研究では、嫌悪性に対しても通常の強化学習を延長した方式を（逆固定型）を使用していた(Millan,1991)。つまり、これまでの研究には嫌悪性に関しては問題がある可能性がある。

次のTable 4-3では、以前にTable 4-2示した報酬性の場合と合わせて、実際シミュレーションに用いた学習の規則を表す。

＜報酬性＞

		$D_R(t-1)$ known		$D_R(t-1)$ unknown
		$D_R(t-1) \neq 0$	$D_R(t-1) = 0$	
$D_R(t)$ known	$D_R(t) < D_R(t-1)$	**SetDist** LearnAct	None	**CreateSvam** LearnAct
	$D_R(t) = D_R(t-1)$	Impossible		
	$D_R(t) > D_R(t-1)$			
$D_R(t)$ unknown				None

＜嫌悪性＞

		$D_P(t-1)$ known		$D_P(t-1)$ unknown
		$D_P(t-1) \neq 0$	$D_P(t-1) = 0$	
$D_P(t)$ known	$D_P(t) < D_P(t-1)$	**SetDist**	None	**CreateSvam**
	$D_P(t) = D_P(t-1)$	None		
	$D_P(t) > D_P(t-1)$	LearnAct		
$D_P(t)$ unknown				None

Table 4-3　完全に確定的な場合に学習を行う条件

確定的環境中における確定的な動作主体の報酬性および嫌悪性ゴールに対する学習規則。

$D_R(t)$, $D_P(t)$:　現時刻の報酬性（嫌悪性）ゴールからの距離。

$D_R(t-1)$, $D_P(t-1)$:　一時刻前の報酬性（嫌悪性）ゴールからの距離。

CreateSvam :　価値評価モジュール内の一つのユニットを有効にし、そのユニットの保持する評価距離を、(現在の評価距離 + 1)に設定する。

SetDist :　(Set Distance)　一時刻前に反応した価値評価モジュール内のユニットの保持している評価距離を、(現在の評価距離 + 1)に書き換える。

LearnAct :　(Learn Action)　直前に行った行動を強化するように出力ニューラルネットワークの学習を行う。（強化信号が出力される）

None :　(None)　何も学習しない。

Impossible :　発生し得ない条件。

4-4-2. 嫌悪性のみのシミュレーション

嫌悪性ゴールに到達する能力の学習による変化を乱数の種の異なるFig.25, Fig.26, Fig.27に表した。ここでもこれまでの報酬性の場合と全く同じ環境を利用し、嫌悪性価値（嫌悪性ゴールまでの距離）学習できるのは１ステップのみとした。嫌悪性ゴールは到達したくない状態であるから報酬性の場合と異なり平均ゴール到着ステップ数を定義することは無意味である。そこで今回はこの指標の代りに１ステップ目のゴール到着確率により動作主体の能力を見積もった。シミュレーションでは嫌悪性ゴールを状態０に割り当て、その状態に直接遷移できる状態は14, 22, 26である（Table 4-1, Fig.4-10参照）。よってそれぞれの上図を見ると学習前のこの確率は(1/31)*(3/4)(=0.24)である。この値は各ゴールへの到達経路に対する回避行動を学習するのにしたがって段階的に(1/31)*(1/4)(=0.08)づつ減少する。この減少に伴ってそれぞれの下図に示す２ステップ以上でのゴール到達確率も次第に減少し、全ての到達経路が塞がれると、その確率は完全に0となる。上のグラフを見ると、まず動作主体がゴールに到達してその回数が増え、それに伴って嫌悪性価値が学習され中間層のユニットの数が一つ増える。次に、この価値に基づいて行動が改善されゴール到達確率が減少する。動作主体が始めてゴールに到達したときには必ず中間ユニットは形成されるが、同じ間違いを二度犯したときにはこの数は増えないので、常にこの数が増加するとは限らない。また価値観が形成されても直ちに正しい回避の行動が身に付くわけではないので、価値の形成に対してゴールに到達する確率が減少するには遅れがある。なお、#3の結果ではまだ２つの経路しか塞がれていない

4-4-2. 報酬性と嫌悪性のシミュレーション

これまで説明してきた報酬性ゴール（状態31）および嫌悪性ゴール（状態0）の両方を持つ環境に対して同様のシミュレーションを行った結果をFig.4-28, Fig.4-29, Fig.4-30に示す。それぞれ(a)は報酬性ゴールに対する、(b) は嫌悪性ゴールに対する動作主体の能力の評価を示している。報酬性ゴールの到達確率は#1,#2の例では報酬性のみの場合（再帰的モデル: Fig.4-18~4.22）よりも最終的な能力が向上したのに対して#3の例では能力が低い。平均ゴール到達ステップ数は報酬性のみの場合と異なり嫌悪性のゴール到達する確率が存在するのでその意味付けはやや不明確になる。よって学習前の動作主体のこのステップ数は報酬性のみの場合よりも小さくなり約17である。#3の例では明かに平均到達ステップ数が増大している部分があるが、これは図(b)と比べてみると嫌悪性ゴールに関する学習が進んだ時に対応する。つまり嫌悪性ゴールに対する回避行動の固定が報酬性ゴールへの到

達を疎外しているのだと考えられる。一方、嫌悪性ゴールに対する評価である図(b)に目を向けると、嫌悪性のみの場合(Fig.4-25~Fig.4-26参照)と異なり、ゴール到達確率（上図）の形状が異なる。まず第一ステップの確率が同じであっても、報酬性ゴールに関する能力の向上に伴って嫌悪性ゴールに到着する確率も減少する。第２にステップの増加に伴う減少がより速くなった。これは報酬性ゴールに到着する確率があるために環境内の確率が速やかに減少することを表している。

この様に、報酬性と嫌悪性を合わせた環境では動作主体の振舞いは複雑になるがこれがより現実に近い状況である。結果としては嫌悪性ゴールを避けつつ報酬性ゴールに到達することができるように適応を行うことができたので成功であると言えよう。

4-5. 確率的な環境に対する適応

4-5-1. 確率的環境と価値観

4-5-1-1. 環境の確率的構造

確率的な環境では同一入力状態で同一の行動をとっても、次の入力状態が一意には決定しない、つまり状態遷移ベクトルが確率的に表現されるのである。しかし、もしもこの遷移が全くランダムに起こるならば環境に知識は存在せず、動作主体はいかなる適応もできないので、環境には確率的な構造が存在する必要があるだろう。この様な不確定性が発生する原因は、環境自体が確率的である単純な場合だけでない。一つには、当然実際的環境では動作主体は全ての情報を取り込む事はできないので隠れた変数が存在する。さらに一般化や推論は確実なものではないから、それらの処理にともない確定性が現れる。これらの不確定要素は実際的にはそう簡単には区別できないし、原理的に区別できない場合も有り得る。環境自体が確率的だとする捉え方は、その他の２つのケースに対してもある程度有効であるから、本論文においては、この確率的な構造を持つ環境を議論の対象とする。

4-5-1-2.シミュレーションにおける環境の動作

シミュレーションに用いた環境は、確率的なので4-2-1.でに説明した式(4-8)に基づいて動作する。ある入力状態においてある行動を選んだときの最も確率の高い遷移ベクトルはこれまで用いてきた環境と同じであるとし、それ以外の状態にはより低い確率で等確率に遷移するものとする。

そこで、ある状態である動作を選んだときの遷移ベクトルの選択確率を

	経路の数	選択される確率	合計
主な経路	1	0.67	0.67
それ以外の経路	31	0.00105	0.33

Table 4-4 環境の選択経路の選択比

として、シミュレーションを行った。つまり、1/3の確率で思いもよらない状態に飛ばされてしまうという、かなり難しい環境である。

4-5-1-3. 価値観の設定方針

確率的な構造を持つ環境では動作主体はより慎重に価値観や行動を組織化する必要があ

り、確定的な環境に適応する場合とは異なった能力が要求される。適応の結果による動作主体の能力は形成された行動と直接関係するが、それ以前に行動生成の指標となる価値体系を適切に形成する必要がある。つまり、ある入力状態に与えられた価値観は引き続き形成される行動に対してスカラーポテンシャルのように作用するので、行動すべき方向に沿って報酬性価値観が増大する必要があるし、嫌悪性価値観は減少する必要がある。

確定的な環境では報酬性のゴールに対するステップ数は最短距離をとればよかったのだが、確率的な場合にはそのステップ数は期待値としてしか見積もれない。仮に、環境の確率構造を完全に把握できるスーパーバイザーが存在しても、ある状態からゴールまでの距離の期待値を計算するには無限ステップの計算を実行する必要があるので、この距離を定義するのは非常に難しい。

さらに、逐次的に経験を増加させる状況では距離と確率の評価の他に信頼性の評価がある。つまり、ある状況におけるある行動とその結果に関わる経験の数が多いほど確率が大きくても小さくてもその確率自体に対する信頼性は向上する。そこで、有効な戦略には確実性も望まれるので、一見手間のかかる方法であっても保守的にこれまでの経路を踏襲するのは意味がある。

この距離、確率および信頼性の３つを別々に表現することも可能であるが、その場合でも行動形成ののためにはこれらを総合した評価を設定する必要があるので、非常に複雑になる。

この様な状況の中でアルゴリズムを論理的に導き出すのはかなり難しい。そこで当面、価値観を構成するモデルを作り、その環境に対する適応能力をを検証し、様々な環境の中で発生する問題点を徐々に改善して行く。

4-5-2. シミュレーションにおける動作主体の動作

4-5-2-1. 価値評価モジュール(VAM)の学習規則

距離と確率を取り扱う規則を検討した。距離が遠くなっても確率が減少しても価値は増加するので、確率の低下を距離の増大で表現するモデルを提案した。

すると、4-2-2-3.では生成規則のみであった価値変更規則は新たに三つ追加された次の四つにまとめられる。

生成規則　　CreateSvam

ある反応すべき入力パターンに反応するユニットがなかったら新たにその入力パターンに反応するユニットを生成し、一時刻前の状

態に対する評価距離を（現在の評価距離＋１）に設定する。

価値増加規則（距離の減少）　**SetDist**

一時刻前に評価距離が与えられ、現在の評価距離がその一時刻前に評価距離より近くであったならば、一時刻前の状態に対する評価距離を（現在の評価距離＋１）に設定する。

価値減少規則（距離の増加）　**ExtDist**

一時刻前に評価距離が与えられ、現在の評価距離がその一時刻前に評価距離と同じかより遠くであったならば、一時刻前の状態に対する評価距離を1step延ばす。

消滅規則　**ExtDist**

価値減少規則により評価距離が遠くになりすぎたら、その予測を行った中間ユニットを消滅させる。

この規則は報酬性と嫌悪性に対して全く同様に適用することができる。嫌悪性に対する価値減少規則の適用は嫌悪性の入力状態からうまく脱出する行動を修得した後には次第にその価値を忘れるような状況である。「天災は忘れた頃にやってくる」という格言はまさにこの状況を表しているだろう。この難点を避けるために嫌悪性価値の減少および忘却させることをやめてしまうこともできるが、すると行動の選択範囲を必要以上に自ら制限する問題が発生する。

価値減少規則と消滅規則では動作タイミングが同じなので、次ページに示すTable 4-5での表示に沿って、ExtDistで統一して表記してある。

ここで二つの評価距離変更規則をより詳しく表すと、

1). 評価距離の更新　**CreateSVAM & SetDist**

$$D[s(t-1)] = D[s(t)] + 1 \qquad (4\text{-}21)$$

2). 評価距離の延長　**ExtDist**

$$D[s(t-1)] = D[s(t-1)] + 1 \qquad (4\text{-}22)$$

ここで、　D[]：入力状態に対応付けられた評価距離
s()：環境から動作主体への入力状態
t：時間

この価値評価モジュールの学習規則はニューラルネットワークでも問題無く実現できるであろう。

4-5-2-2. 出力ニューラルネットワークの(ONN)の学習規則

確率的環境ではゴールに到達できても、偶発的な出来事であって必ずしも適切な行動をとった結果ではないかも知れないので、後の経験の中で望ましい結果が得られなければ行動を変更する必要がある。つまり、動作主体は一度成功したからといって、盲目的に同じ動作を繰り返すのではなく、時には行動を変えてみたり場合によってはある動作を諦めることが必要である。すると、時により有効なゴール到達ルートを発見することも期待できる。

行動学習による出力ニューラルネットワーク内の確率行列の学習は4-4.で説明したように報酬性と嫌悪性の場合で異なる。4-2-2.で説明した、式(4-10)における確率行列Wの行列要素をW_{ij}としてその変更規則を示す。

ここで、 $D_i[]$: 入力状態に対応付けられた評価距離
(i = R, P : 報酬性, 嫌悪性)
s() : 環境から動作主体への入力状態
t : 時間
W_{ij} : 動作確率行列の行列要素 (i = 出力状態, j = 入力状態)

とすると。

報酬性価値観の学習条件

$$\text{if} \begin{cases} D_R[s(t)] - D_R[s(t-1)] < 0 \\ \text{or} \\ D_R[s(t)] \text{ known and } D_R[s(t-1)] \text{ unknown} \end{cases} \tag{4-23}$$

嫌悪性価値観の学習条件

$$\text{if} \begin{cases} D_P[s(t)] - D_P[s(t-1)] > 0 \\ \text{or} \\ D_P[s(t)] \text{ unknown and } D_P[s(t-1)] \text{ known} \end{cases} \tag{4-24}$$

行列の変更規則

$$W_{ij} = \begin{cases} W_{ij} + \dfrac{(1 - W_{ij})}{2} & (i = s(t-1)) \\ \dfrac{W_{ij}}{2} & (\text{else}) \end{cases} \tag{4-25}$$

以上の学習規則をあわせて次ページのTable 4-5に表す。ただし上記のゴールまでの距離とTable 4-5内の距離表示は

$$D_i(t) = D_i[s(t)] \qquad (i = R, P : \text{報酬性, 嫌悪性})$$

の関係をもつ。

なお、報酬性と嫌悪性の両方を同時に取り扱うと学習方針が衝突することがある、これは今後検討すべき重要な問題ではあるが、本論文では嫌悪性による制御を優先させることとしている。

＜報酬性＞

		$D_R(t-1)$ known		$D_R(t-1)$ unknown
		$D_R(t-1) \neq 0$	$D_R(t-1) = 0$	
$D_R(t)$ known	$D_R(t) < D_R(t-1)$	**SetDist** LearnAct	None	**CreateSvam** LearnAct
	$D_R(t) = D_R(t-1)$	**ExtDist**		
	$D_R(t) > D_R(t-1)$			
$D_R(t)$ unknown				None

＜嫌悪性＞

		$D_P(t-1)$ known		$D_P(t-1)$ unknown
		$D_P(t-1) \neq 0$	$D_P(t-1) = 0$	
$D_P(t)$ known	$D_P(t) < D_P(t-1)$	**SetDist**	None	**CreateSvam**
	$D_P(t) = D_P(t-1)$	**ExtDist**		
	$D_P(t) > D_P(t-1)$	**ExtDist** LearnAct		
$D_P(t)$ unknown				None

Table 4-5　一般的な場合に学習を行う条件

確率的環境中における確率的な動作主体の報酬性および嫌悪性ゴールに対する学習規則。

$D_R(t)$, $D_P(t)$:　現時刻の報酬性ゴールからの距離。

$D_R(t-1)$, $D_P(t-1)$:　一時刻前の報酬性ゴールからの距離。

CreateSvam : (Create SVAM) 価値評価モジュール内の一つのユニットを有効にし、そのユニットの保持する評価距離を、(現在の評価距離 + 1) に設定する。

SetDist : (Set Distance)　一時刻前に反応した価値評価モジュール内のユニットの保持している評価距離を、(現在の評価距離 + 1)に書き換える。

ExtDist : (Extend Distance) 一時刻前に反応した価値評価モジュール内のユニットの保持している評価距離を一つ長くする。また、評価距離が大きくなりすぎたら、そのユニットを無効にする。

LearnAct : (Learn Action)　直前に行った行動を強化するように出力ニューラルネットワークの学習を行う。（強化信号が出力される）

None : (None)　何も学習しない。

4-5-3. シミュレーション結果

確率的環境における慎重に動作を学習する動作主体の有効性を確かめるために、三種類の条件でシミュレーションを行なった。動作主体は慎重な場合とそうでない場合について分類し、慎重でない動作主体の行動変更(ONN)は一度成功した動作の選択確率を直ちに1とする従来の手法を用い、慎重な動作主体は一度成功しても完全には行動を確定せず成功を繰り返すにつれて次第にその確率を高める手法を用いた。なお、価値観(VAM)に関しては一度決めてからも取り消すことができる前節のアルゴリズムをを用いた。環境は確定的な場合と確率的な場合について分類し、確定的な環境は前節までと同じもので、確率的な環境は4-5-1-2. で説明したように、主な遷移経路は従来どおりとしつつそれ以外の状態にも遷移する確率構造を持つものとした。

はじめに慎重な動作主体が従来どおりの確定的な環境において適切に動作するかを確認し、その能力を確定的な動作を行なう動作主体と比較した。次に従来どおりの慎重でない動作主体が確率的な環境において、どの様な形で適応の困難をあらわするかを示した。そして、最後に慎重な動作主体は確率的な環境においてもより確実に適応できることを示した。

4-5-3-1. 確定的な環境における慎重な動作主体

環境が確率的であるにも関わらず動作主体が慎重な場合のシミュレーションの例をFig.4-31, Fig.4-32に示した。なお今後調べる確率的環境ではゴールに到達するのは難しくなるのでゴール到達確率の対数メモリのスケールを変更したが、Fig.4-31(b)はこれまでのシミュレーション結果との比較のために従来どおりのスケールでFig.4-31(a)の上図と同じグラフを再表示した。

動作主体が確率的に慎重に行動の選択を行なったならば、確定的な環境に対しては最終的な能力は同じとなるはずだが、慎重である必要のない環境で慎重になることは学習速度の低下をまねく。動作決定のための乱数の種として同じ初期値を用いたFig.4-18Fig.4-31及びFig.4-19Fig.4-32の結果を比べると、やはり後者の方が能力の向上に若干時間を要する。その他の振舞いは再帰的なモデルの場合とほぼ同様で平均ゴール到達ステップ数は3.5から4程度になる。なお、第一ステップでゴールに到達する確率はこれまでの結果と異なり、徐々に増加する。

4-5-3-1. 確率的な環境における慎重でない動作主体

環境を確率的としたシミュレーション結果をFig.4-33, Fig.4-34, Fig.4-35に示す。Fig.33

の上図を見ると、一ステップでゴールに到達する確率が４つの段階を経て増加しているので、ゴールに直接到達する４つの経路が全て学習されたことを示している。しかしながらこれらの遷移は確率的なのでこれまでの確定的な環境の場合に比べてはじめの１ステップで到達する確率が小さくなっている。後の二つの結果では４つの段階は現れていなが、最終的な値から４つの経路を学習していることがわかる。

慎重でない動作主体が確率的な環境で動作しているので、はじめに学習した行動が偶然適切であった場合には能力が向上するが、そうでない場合には不適切な行動を覚えてしまうので返って能力が低下する。よって、そのゴール到達能力は経験に大きく依存する。ここに示した３つの例では#2の結果はかなりうまく適応し、平均ゴール到達ステップ数は7.5程度まで小さくなったが、#3ではあまり適応に成功せずその値は約9であり、さらに#1の結果では平均ゴール到達ステップ数は800ステップ後においても17.5程度である。#1の例では400ステップ目頃に価値評価モジュール内の中間ユニットの数が明かに減少している。これは、それ以前の早い段階で不適切に生成されたユニットが、後に価値減少規則と消滅規則により消滅したことに起因する。また80ステップ目や200ステップ目付近で見られる平均ゴール到達ステップ数の増大は不適切な行動を覚えたことを示している。なお、環境の変化にともない学習前の平均到達ステップ数は約29に減少している。

このように、確率的な環境における慎重でない動作主体は時に不適切な行動を覚えてしまい、そのゴール到達能力が悪化すると言う問題点がある。この問題を解決した慎重に動作を行なう動作主体のシミュレーション結果を次に示す。

4-5-3-2. 確率的な環境における慎重な動作主体

確率的な環境における慎重な行動学習を行なう動作主体の結果Fig.36, Fig.3-37, Fig.3-38を見ると、慎重でない動作主体と比較して明かにゴールに到達する能力が向上している。まず平均ゴール到達ステップ数はそれぞれ対応する慎重でない場合に比べて減少している。最も効果があったのは#1の場合で、慎重でなかった場合の17.5から半分以下の7程度に減少している。結果的に3つの例の平均到達距離のバラツキは小さくなり、その値は6から7程度となる。つまり、慎重でない場合に比べてより確実にゴールに到達する行動を組織化することができることを表している。

これら３種類のシミュレーションの結果からより一般的な確率的な環境に対してもこれまでの価値観と行動の学習則を改良することにより適応できることが示された。

4-6. まとめ

第2章で提案された価値観を持つ知能システムの能力を計算機シミュレーションで検証した。本章では、必要とされる出力ニューラルネットワーク、価値評価モジュール及び認識連合モジュールのうち前者二つを含むシステムを研究対象とした。そして初期のいくつかの試みの結果、初期段階の研究においては一般化を含む問題は価値観の形成に関わる問題から分離して議論する方が良いと考えたので、動作主体が適応すべき環境を一般化を行なうことが不可能な迷路とした。そして、まず第2章で示した基本モデル、一次モデル、再帰モデルの3種類の価値評価モジュールをもつ動作主体に対して報酬性ゴールを目指すシミュレーションを行ない、再帰モデルの環境に対する適応能力の高さ、さらに個性豊かな振舞いをすることを示した。次に、嫌悪性価値観に対してもこのシステムが有効に動作することを確認した。そして、最後に動作主体にとってより難しい確率構造を持つ迷路においても、このモデルを改良することにより報酬性ゴールに到達する能力を高めることができることが示された。

このようにして、ゴールに対応する基本的な報酬性と嫌悪性の価値観から次第に価値体系が形成され、さらにされに基づいて適切な行動が形成されることが計算機シミュレーションにより明らかになった。

ここで残された課題は、一つには嫌悪性価値に対する3種類の行動学習のうち本当に脱出記憶型が妥当であったのか、さらにこの3種類以外のアルゴリズムはあるのだろうかという疑問がある。さらに、確率的な環境における距離、確率、信頼性に対する一般的な取り扱いはどうあるべきか。3つ目には報酬性と嫌悪性の両方の価値から行動学習に結び付ける適切な制御方法とは何かという疑問である。これらは今後の課題であろう。

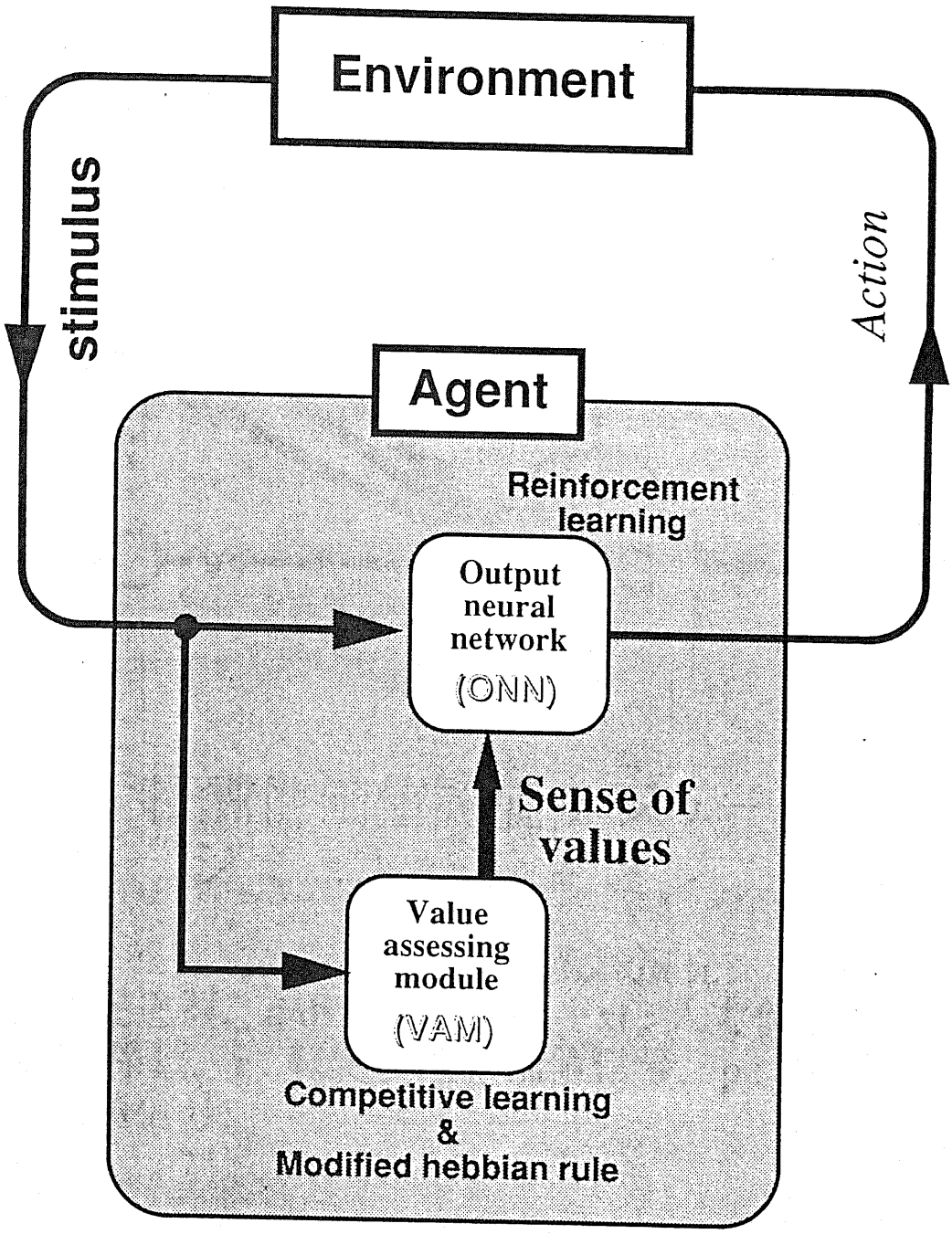


Fig.4-1 環境と動作主体を含めたシミュレーション

出力ニューラルネットワーク(ONN)と価値評価モジュール(VAM)により構成される動作主体(Agent)と環境(Environment)が刺激(Stimulus)と行動(Action)を通して相互作用を行う。

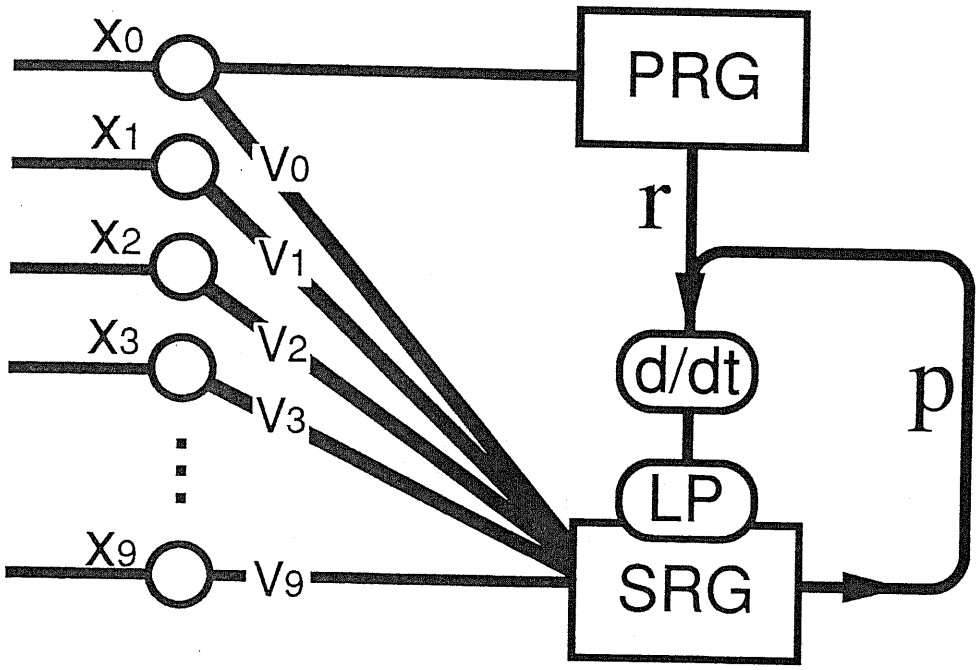


Fig.4-2 報酬性価値評価モジュールのシミュレーション

PRG : 基本報酬発生器, SRG : 二次報酬発生器, LP : 学習促進器, d/dt : 微分回路, xi : 入力信号, Vi : 結合係数, r : 報酬性基本価値信号, p : 報酬性二次価値信号.

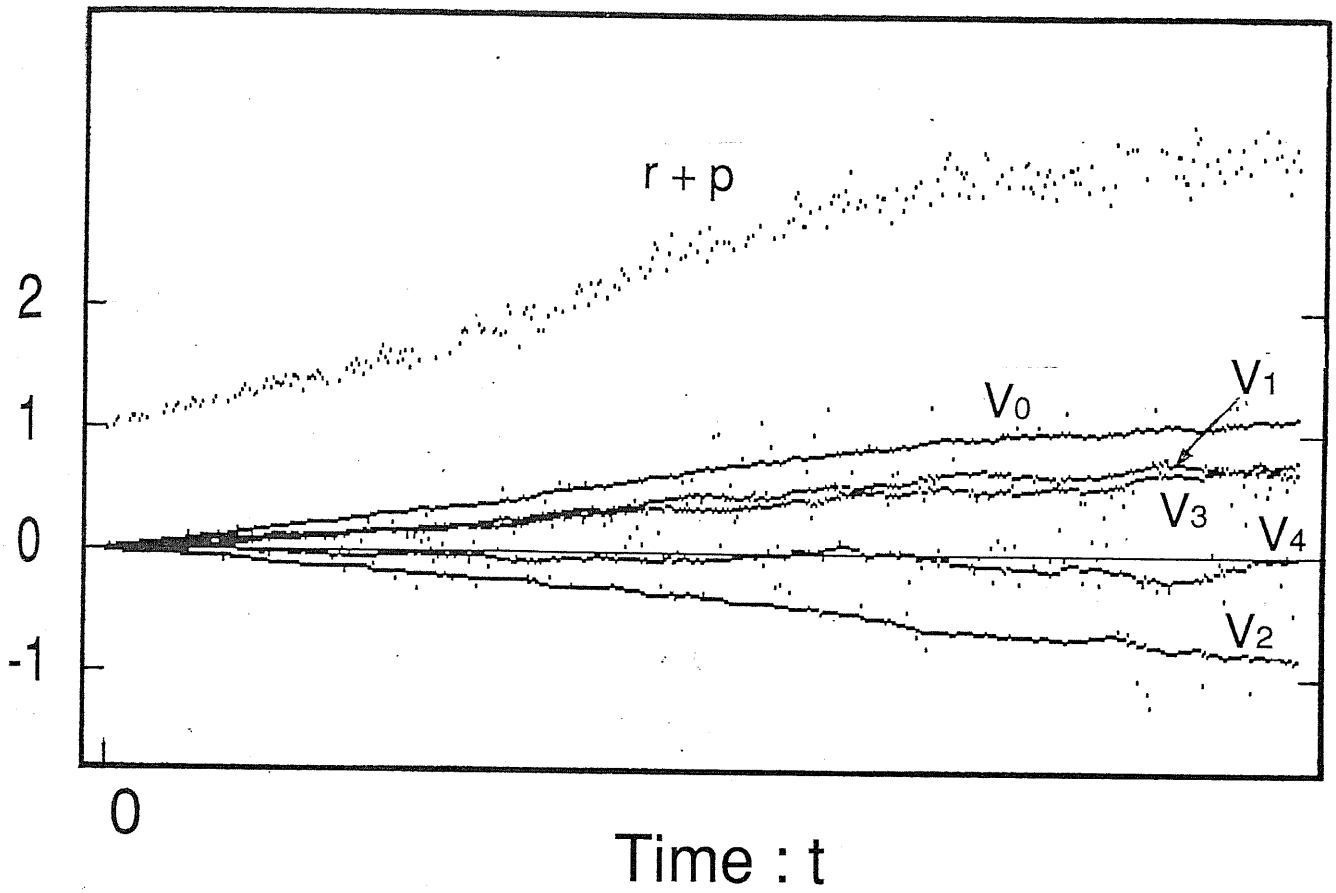


Fig.4-3 報酬性価値評価モジュールのシミュレーション結果

ユニットコーディングの同時刻相関に対するシミュレーション結果。

x_i : 入力信号, V_i : 結合係数, r : 報酬性基本価値信号, p : 報酬性二次価値信号.

x_0が基本価値信号に対応する。x_1とx_3は同時刻の正の相関を持つので結合係数V_1とV_3は増加し、x_2は負の相関を持つので結合係数V_2は減少する。その他のランダムな入力に対する結合はほぼ0となる（グラフではV_4のみ示してある）。なお、これらの値はほぼ一定値に安定する。

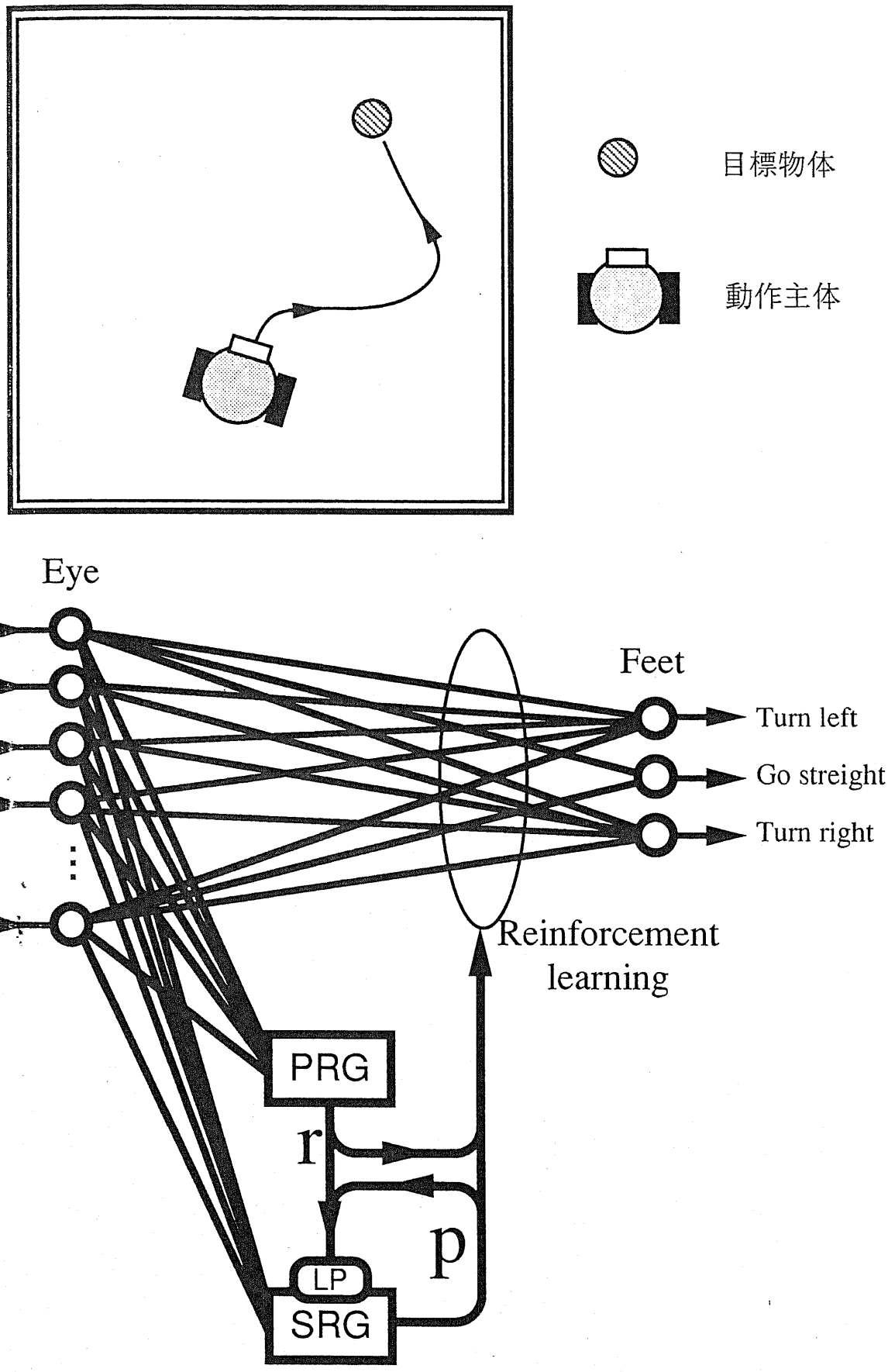


Fig.4-4 二次元空間内で目標物体を捕らえるシミュレーション

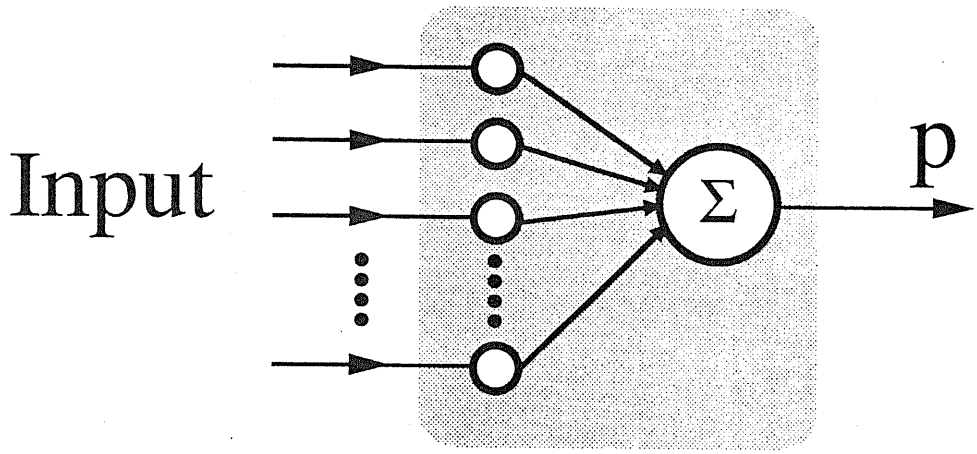


Fig.4-5 二次報酬発生器の構成

第一層は特定の入力パターンに選択的に反応するように学習が行われる。第2層はこれらの出力を集めて二次価値信号(p)を生成する。

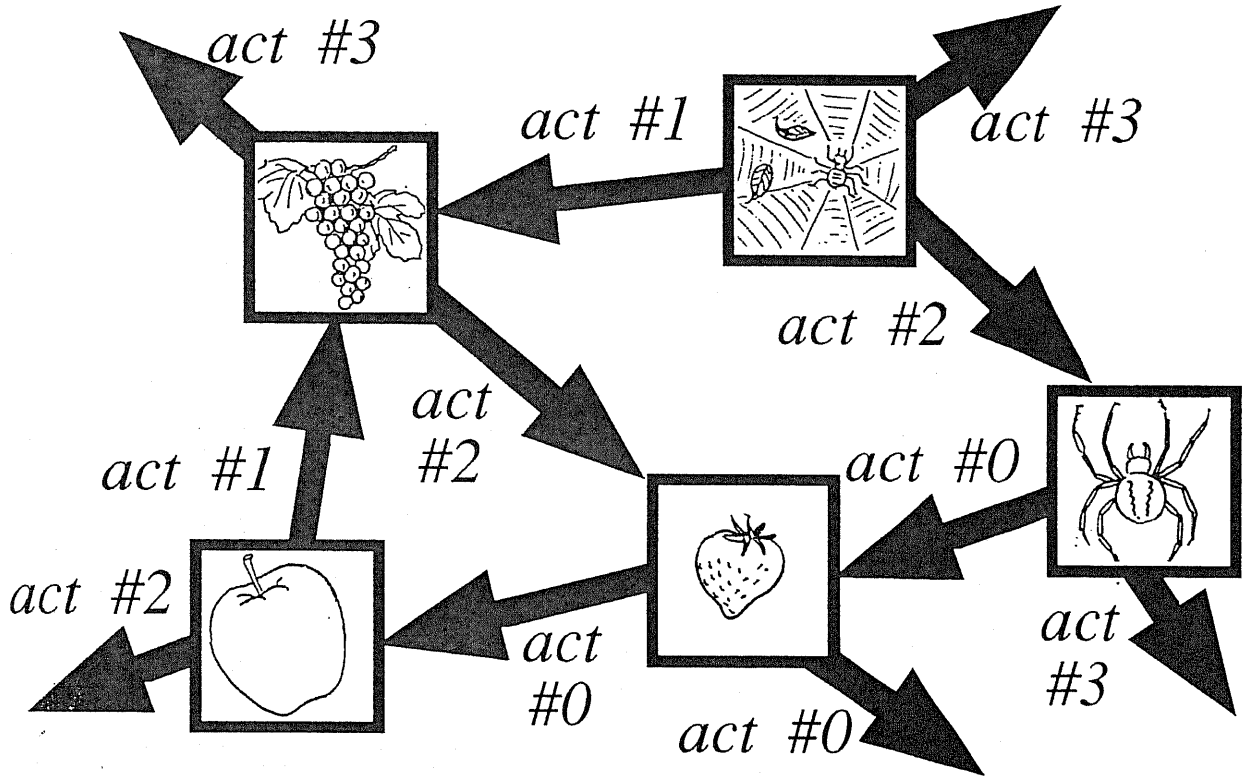


Fig.4-6 マス目と遷移ベクトルによる環境

現在のマス目と行動の選択により次に選ばれる状態が決定する。

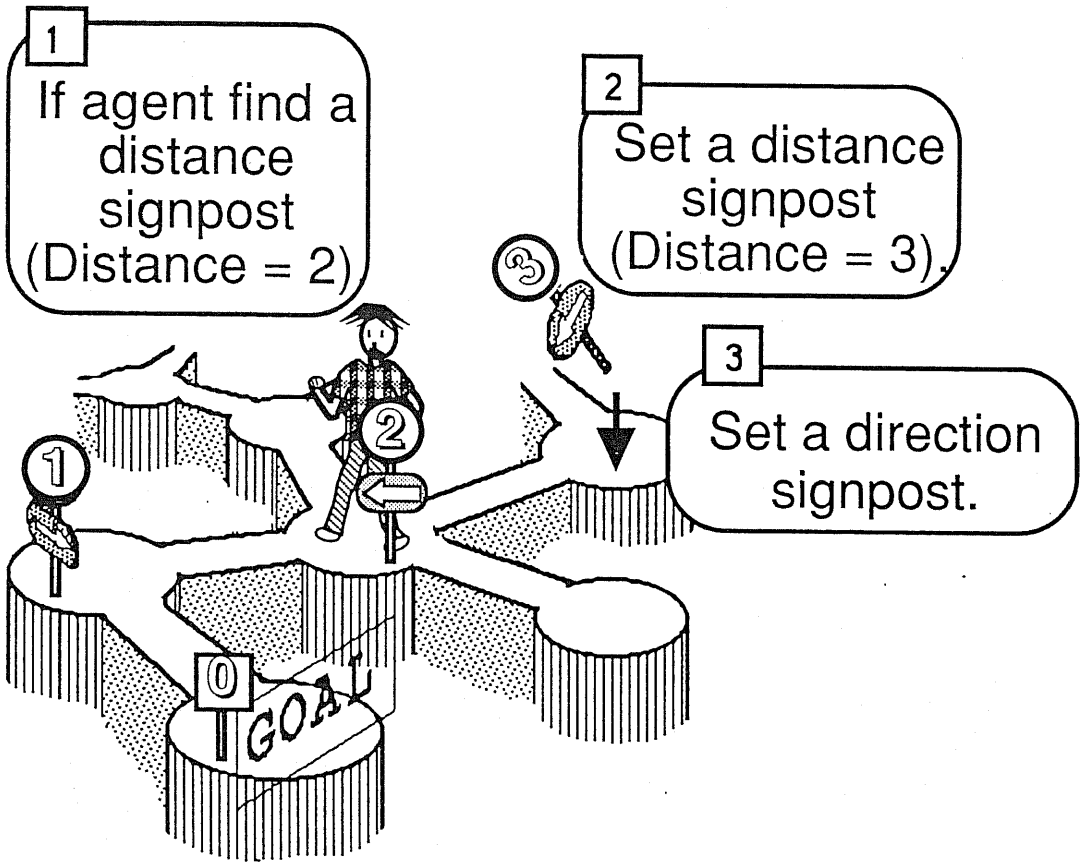


Fig.4-7 迷路の中で標識を立てる動作主体

動作主体が距離0のゴールに効率的に到着できるように距離標識〔○に数字を書き込んである〕と方向標識〔矢印による表示〕を立てて行く様子を描いた。

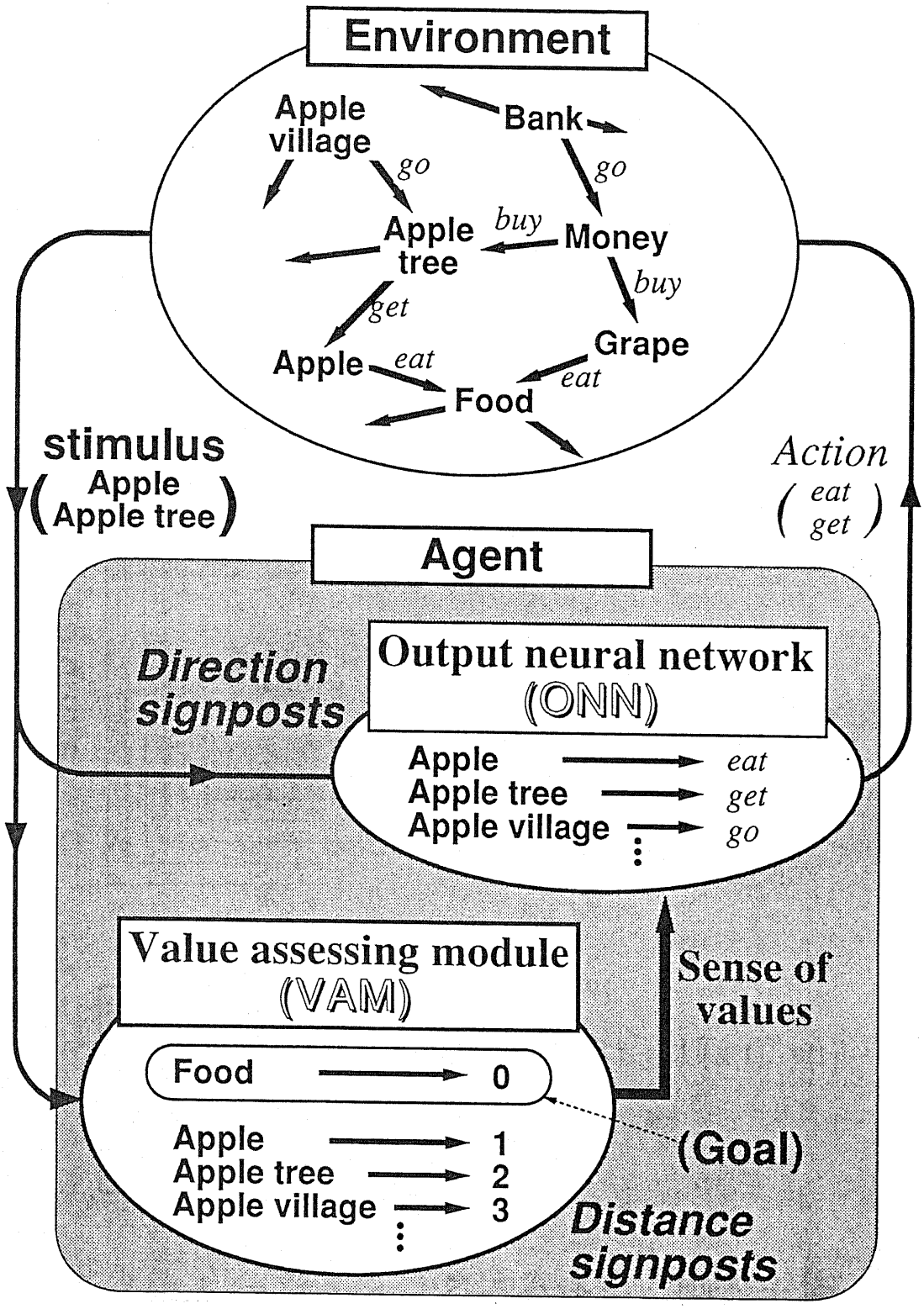


Fig.4-8 ニューロシステムと標識の関係

環境(Environment)と動作主体(Agent)が相互作用を行いながら学習を行う。出力ニューラルネットワーク(ONN)には方向標識に対応するモジュールで、刺激に対する適切な動作が示される。価値評価モジュール(VAM)は距離標識に対応し、刺激に対するゴールまでの距離評価が示される。初期状態においては、ゴールにあたる栄養物(Food)に対する距離が0であることのみを知っている。

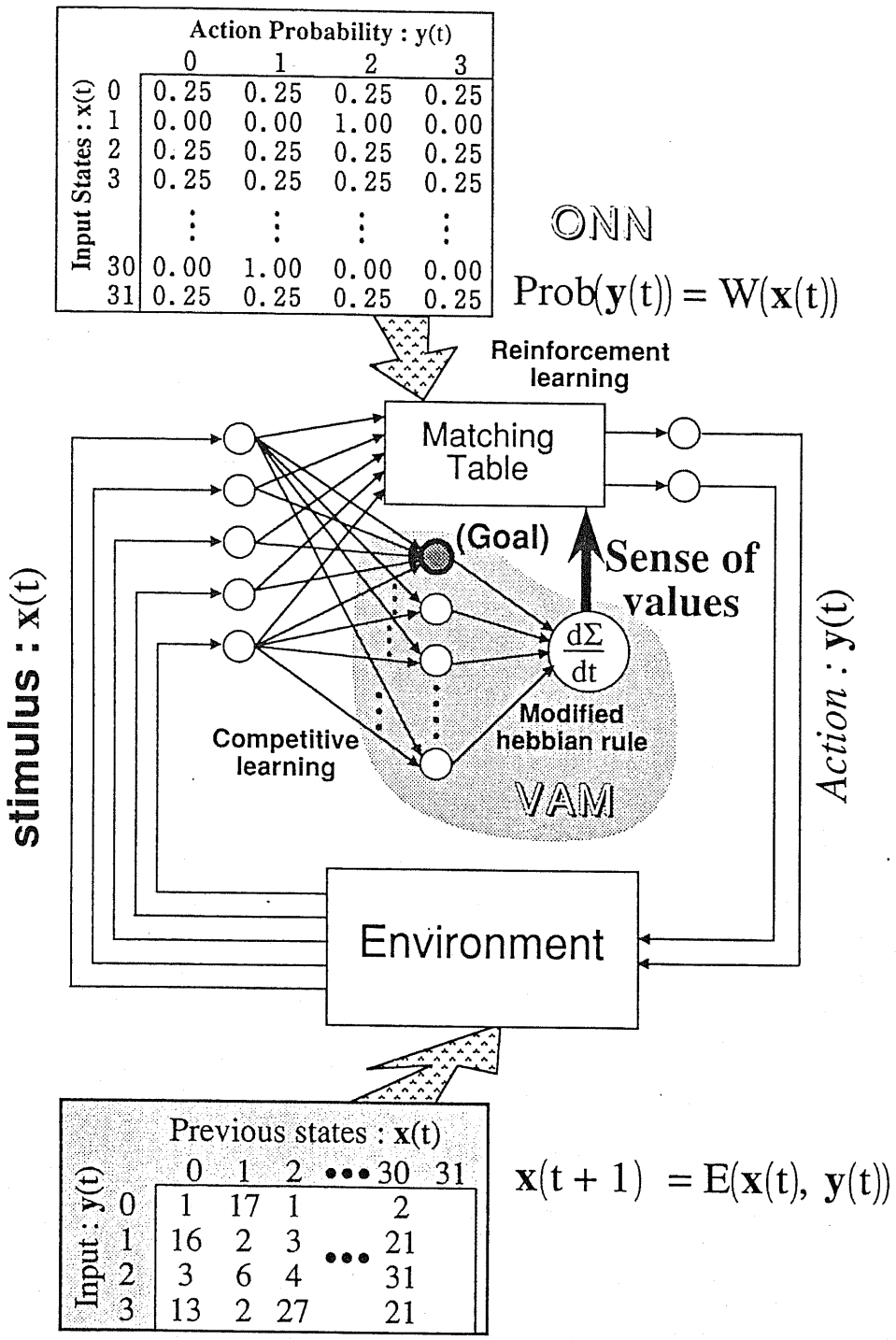


Fig.4-9 シミュレーションにおける迷路と動作主体

環境(Environment)と動作主体(Agent)が相互作用を行いながら学習を行う。環境は確定的に変化する状態遷移行列によって記述される。出力ニューラルネットワーク(ONN)にはマッチングテーブルが使用されている。価値評価モジュール(VAM)は二層構造のニューラルネットワークで、第1層の一番上のユニットが基本価値をコーディングしている。

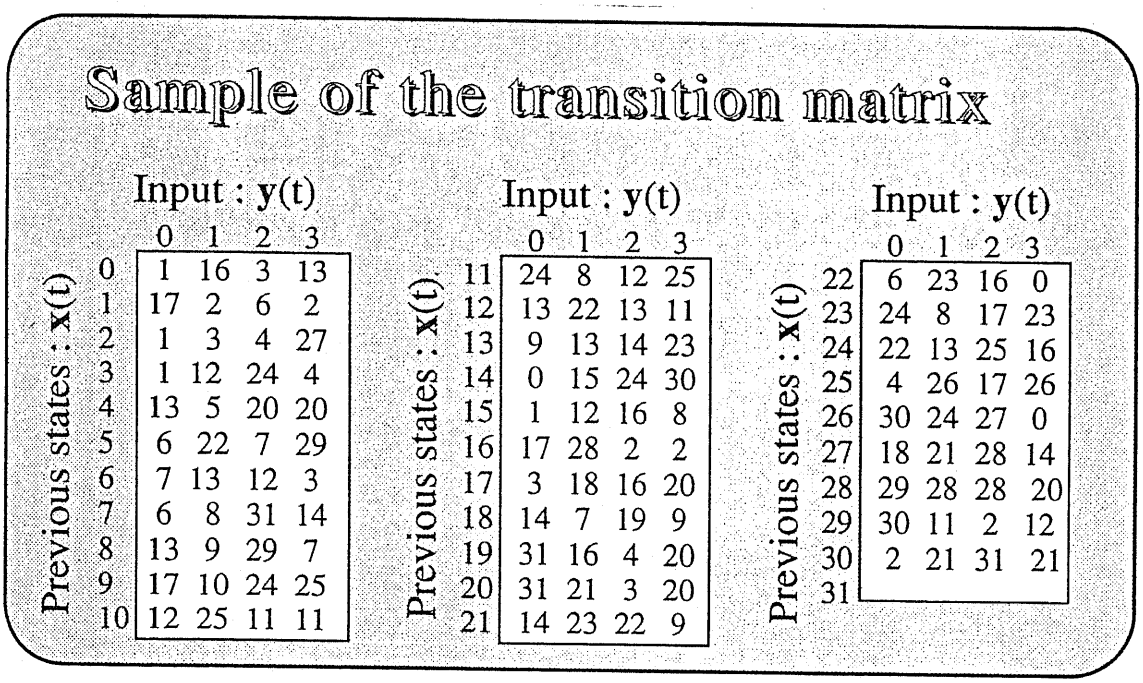
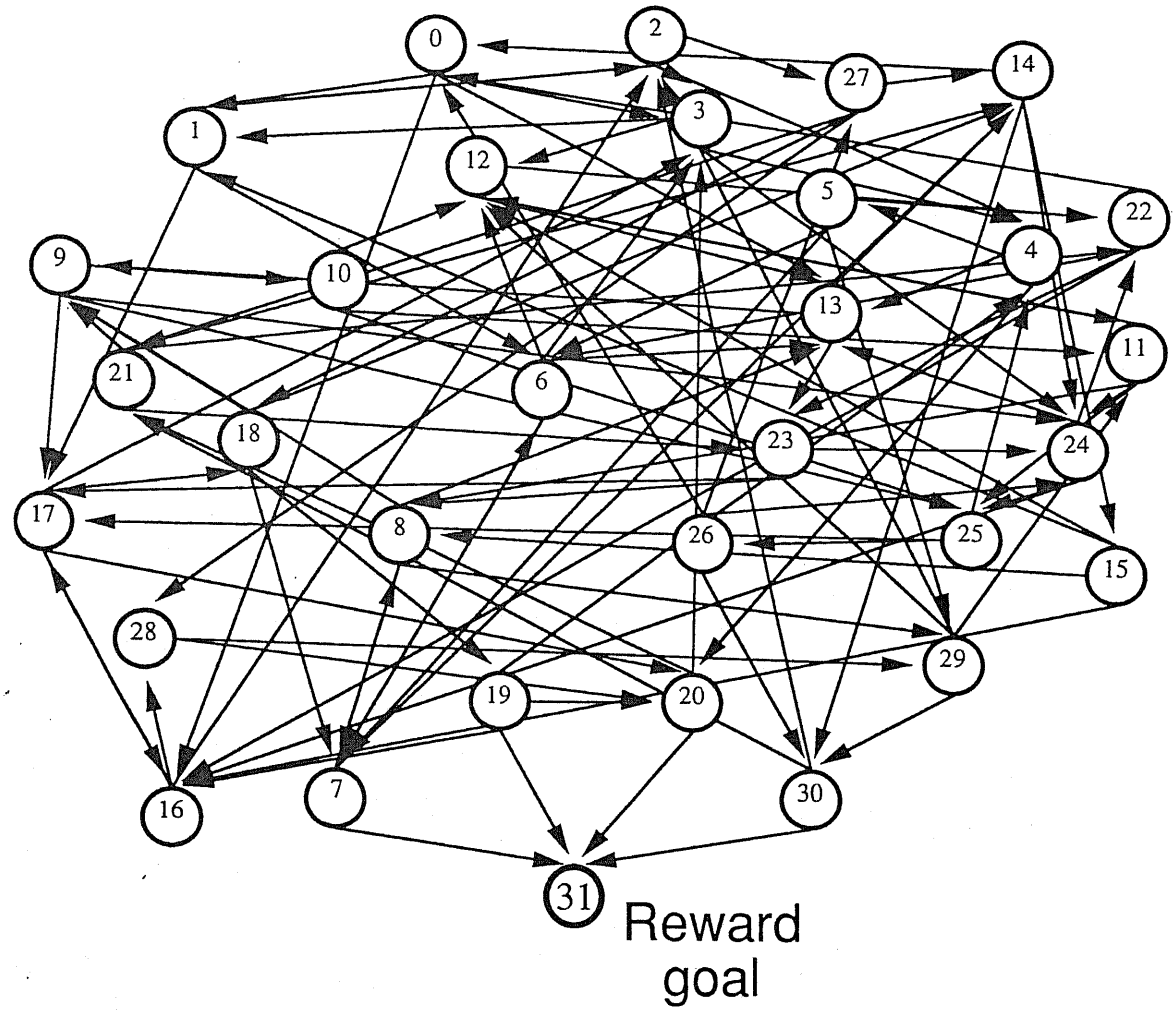


Table 4-1 シミュレーションで用いた環境の状態遷移行列

例えば、以前の状態Previous state)が２で入力(Input)が１だったならば現在の状態は３になる。



Except for the reward goal (31), all states have 4 outgoing arrow.

Fig. 4-10 環境の状態遷移ダイアグラム

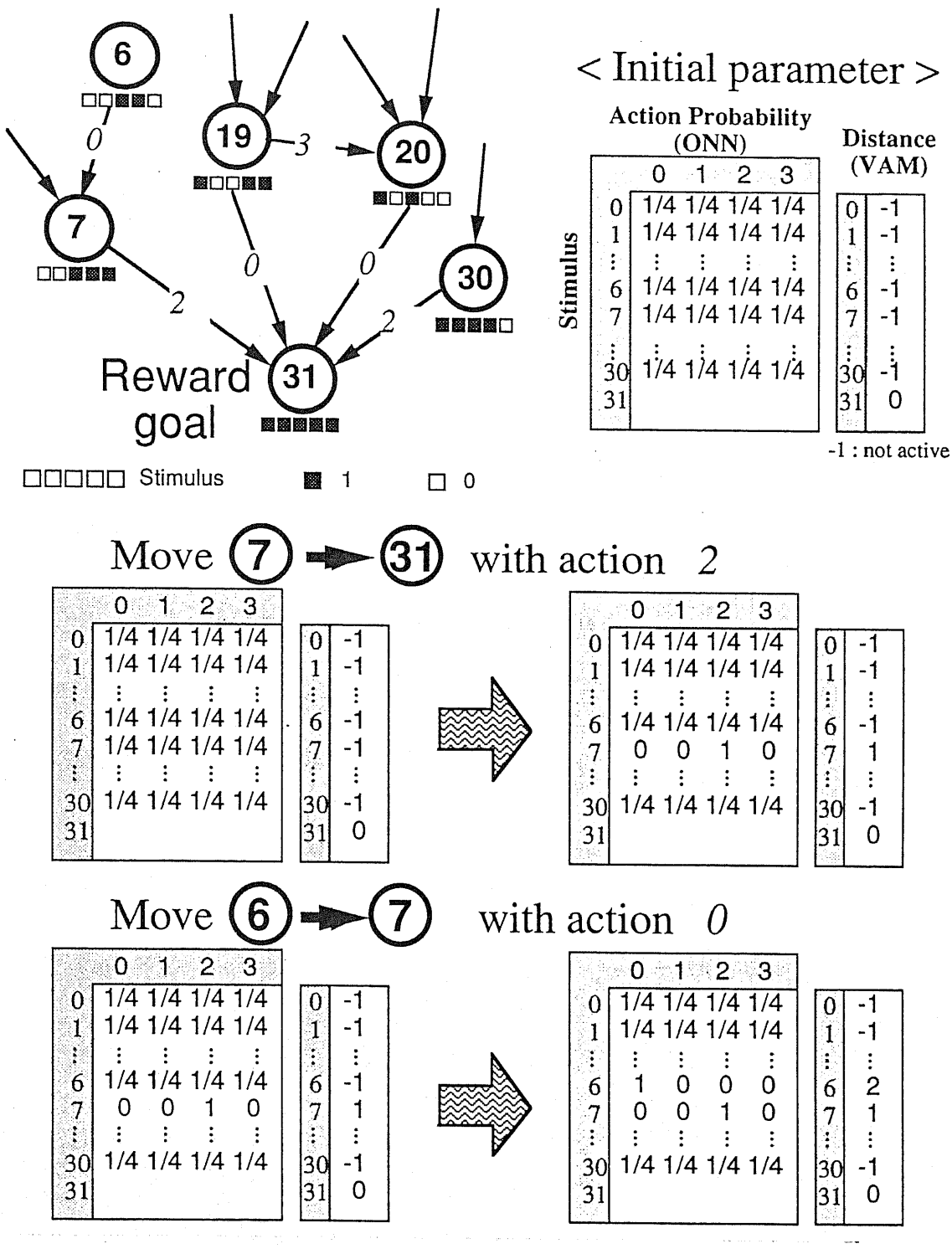
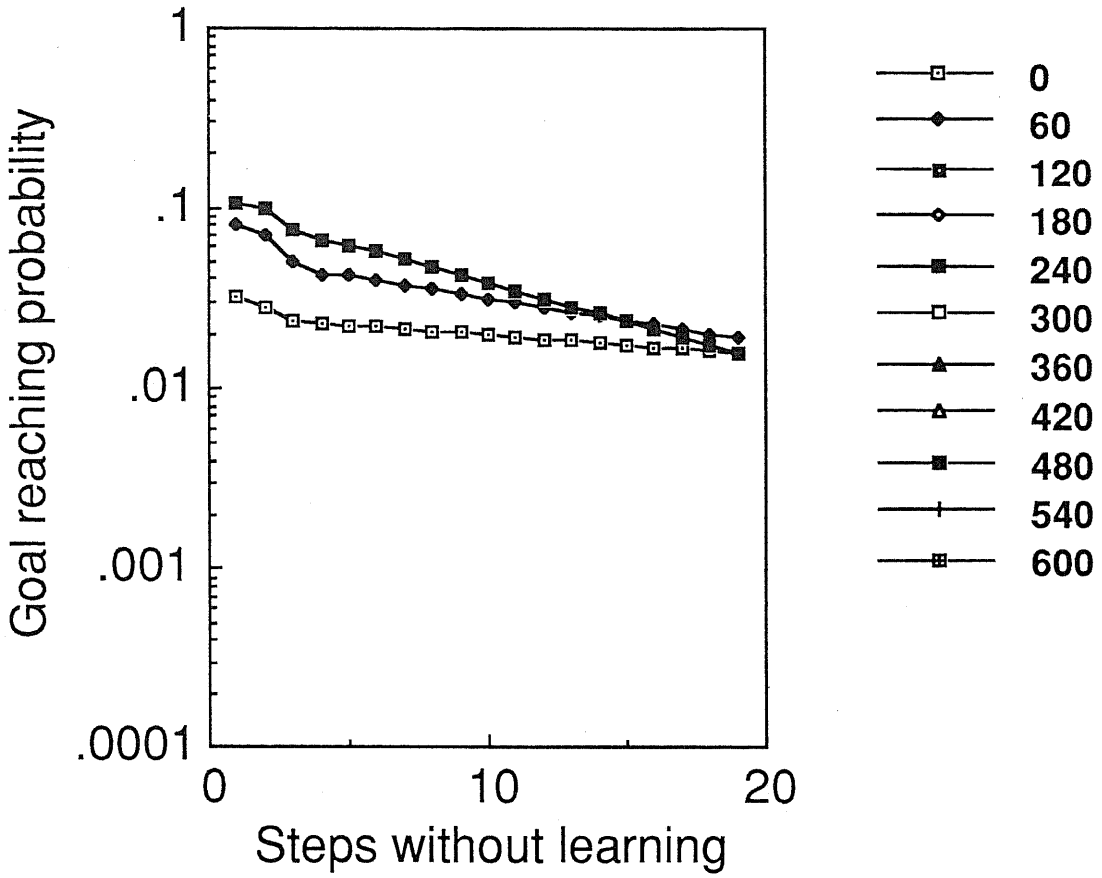


Fig. 4-11 動作主体の学習過程

左上に報酬性ゴール周辺の環境の状態遷移ダイアグラムを示し、その右に動作主体の初期状態の内部パラメータを示した。中段および下段は動作主体が環境中を動きまわってうまく学習することができた場合に、内部パラメータがいかに変化するかを示した。



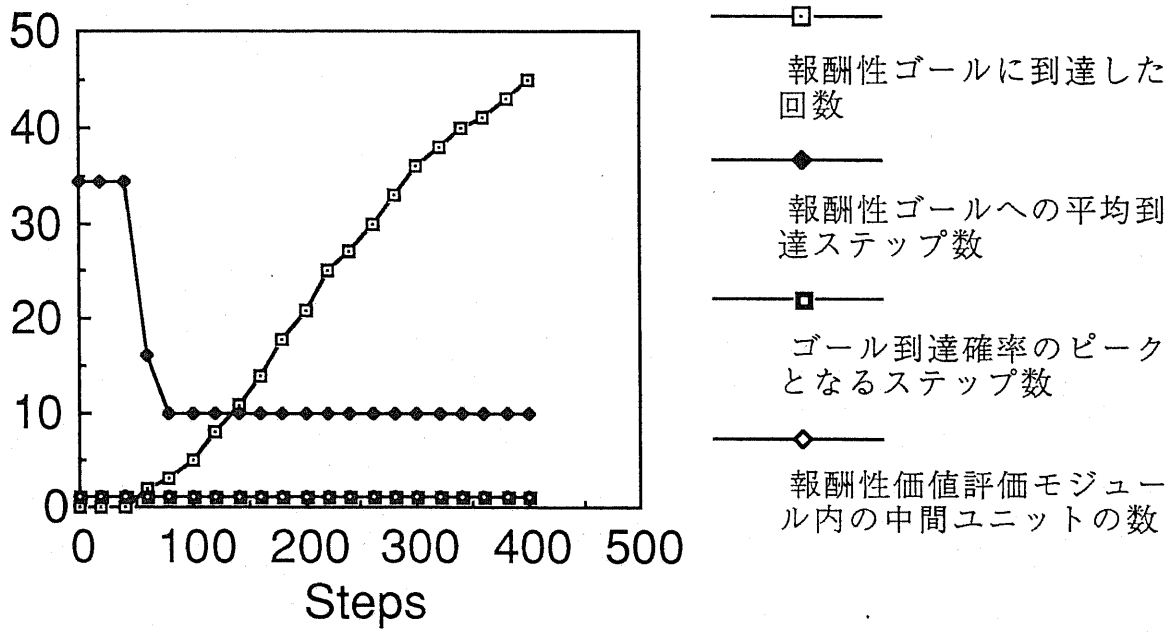
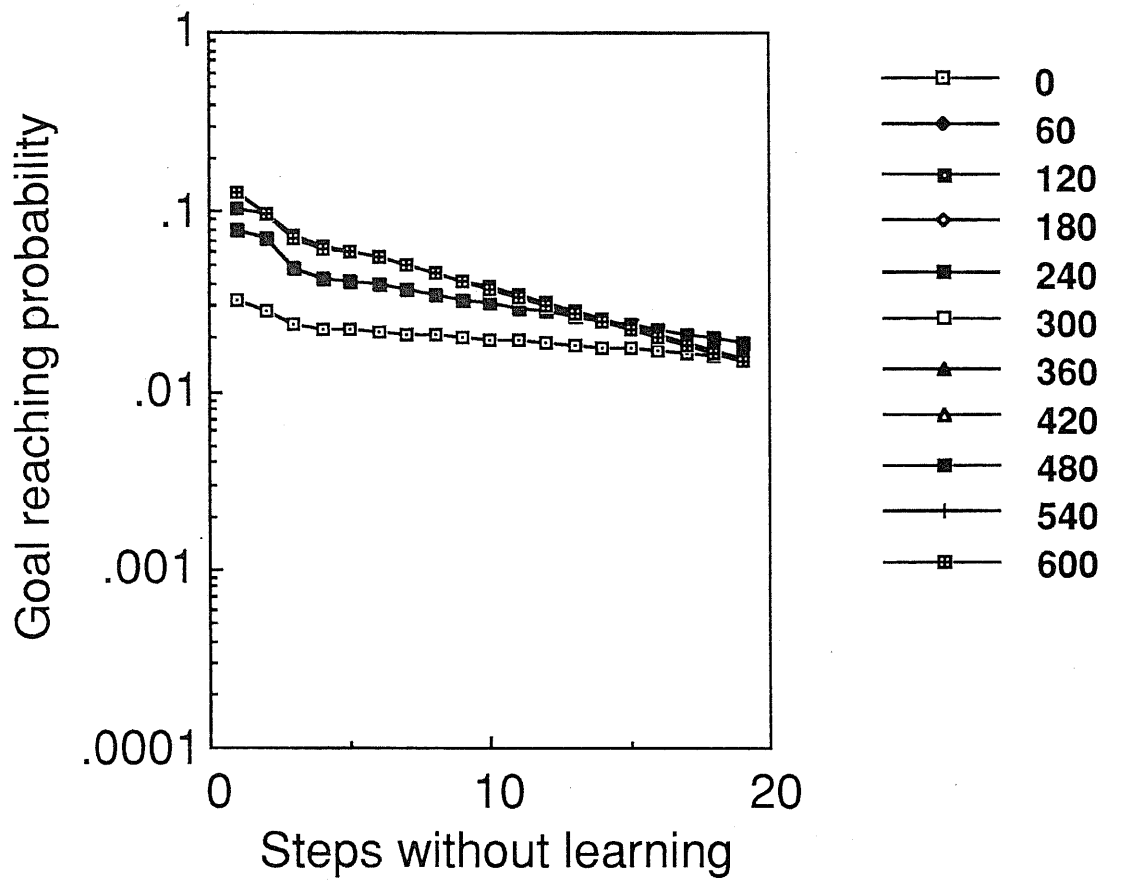


Fig. 4-12 基本モデルのシミュレーション (#1)

上図：学習を止めた状態で報酬性ゴールに到達する確率。
下図：学習に伴う動作主体の状態変化など。



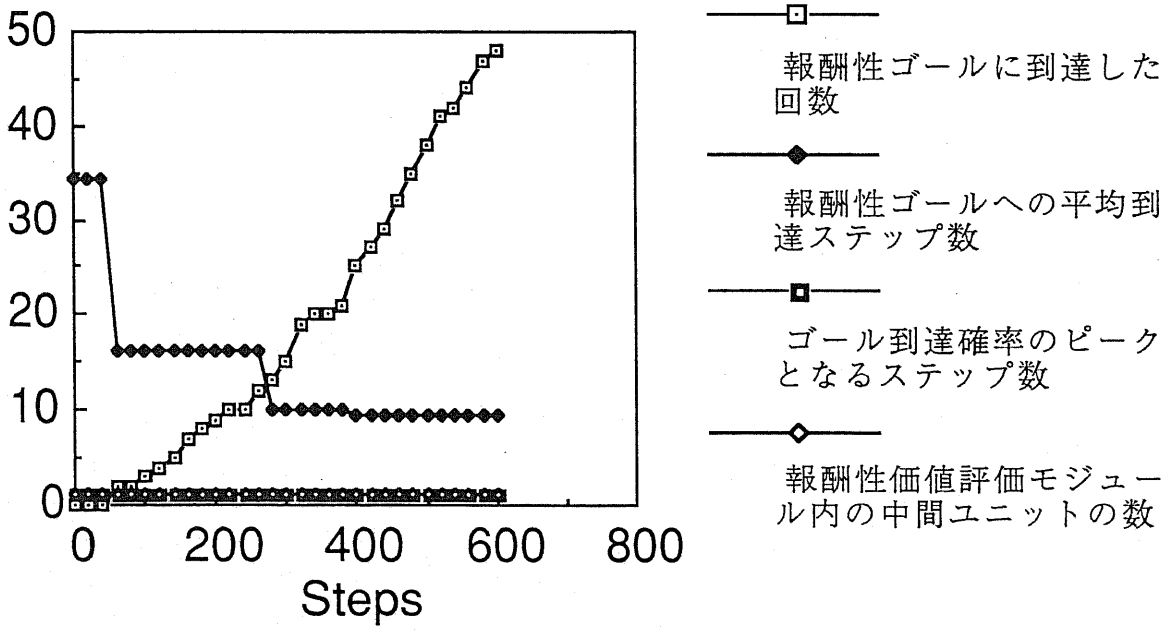
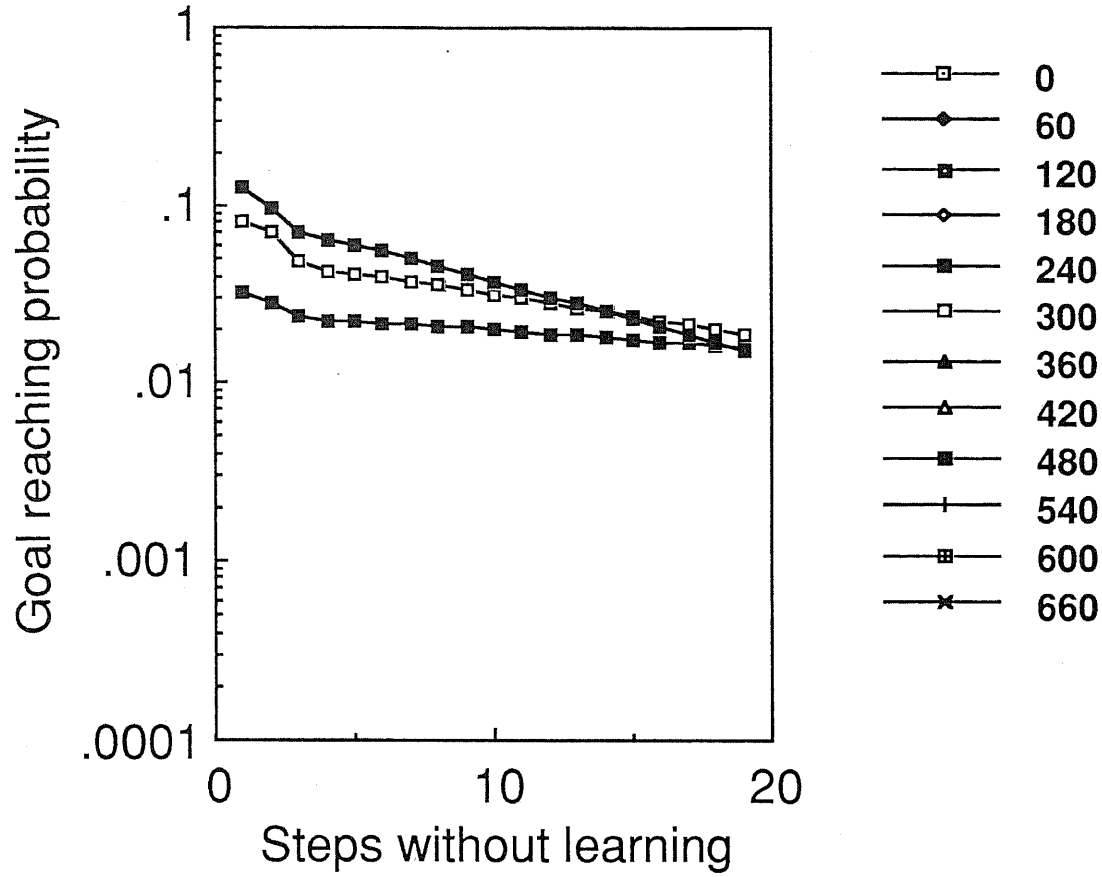


Fig. 4-13 基本モデルのシミュレーション (#2)

上図：学習を止めた状態で報酬性ゴールに到達する確率。
下図：学習に伴う動作主体の状態変化など。



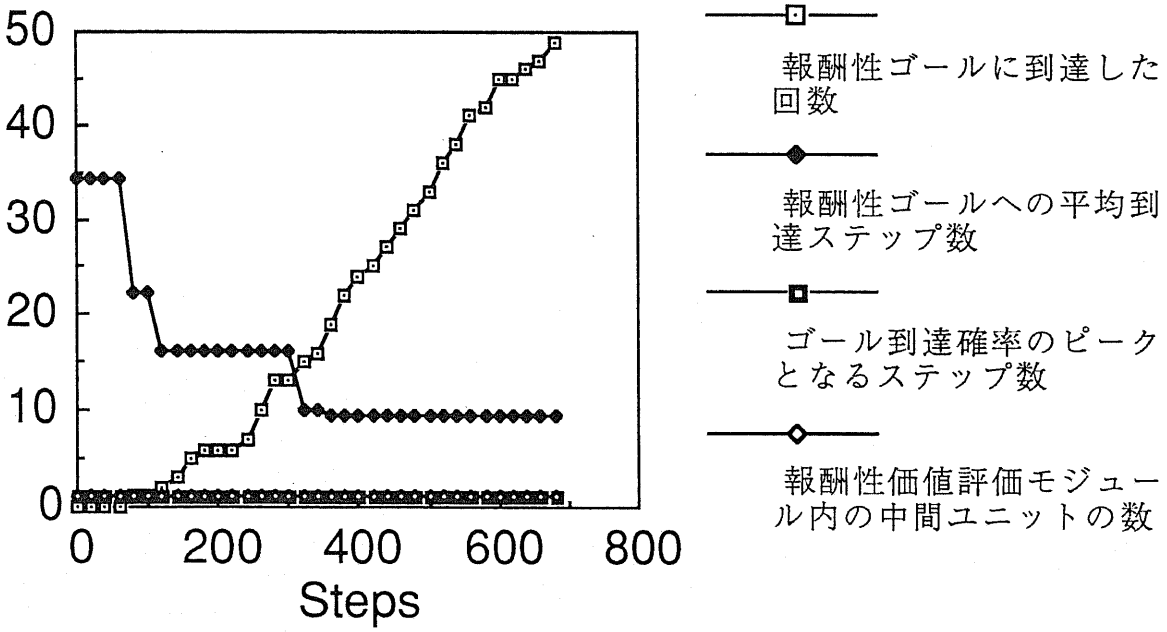
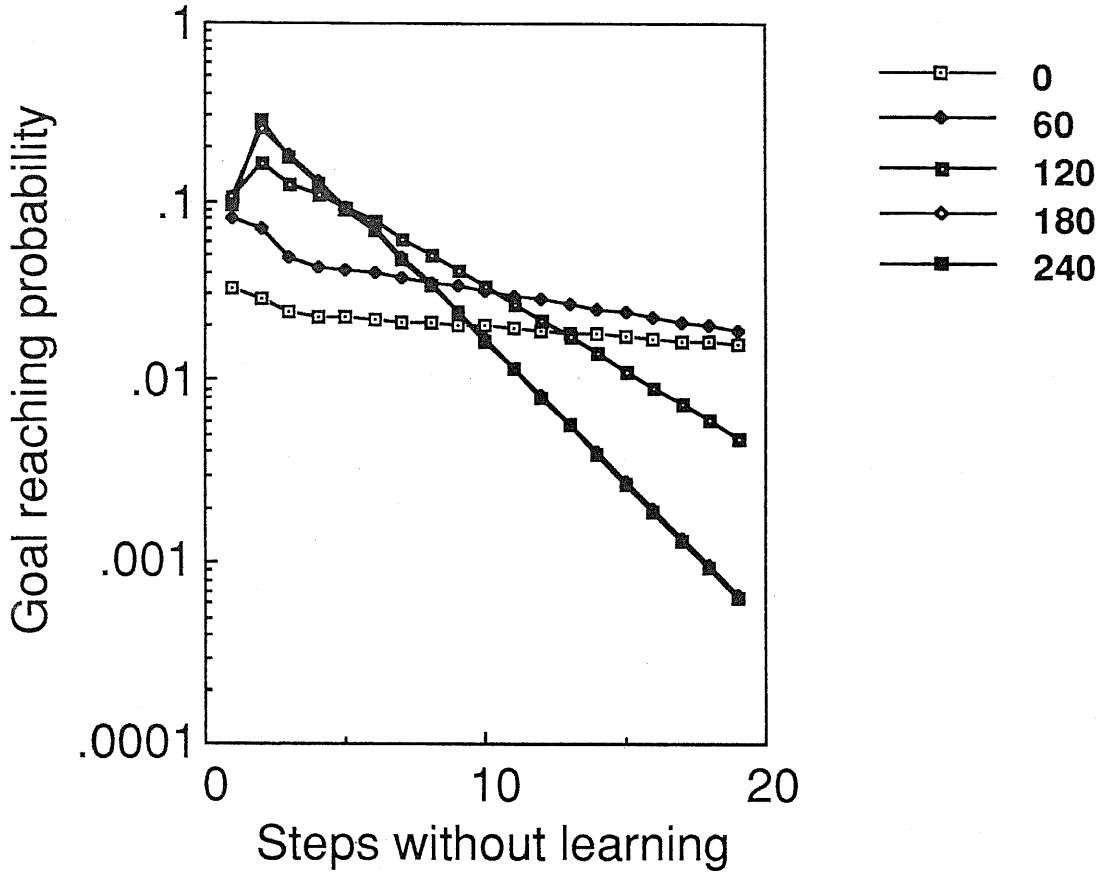


Fig. 4-14 基本モデルのシミュレーション (#3)

上図：学習を止めた状態で報酬性ゴールに到達する確率。
下図：学習に伴う動作主体の状態変化など。



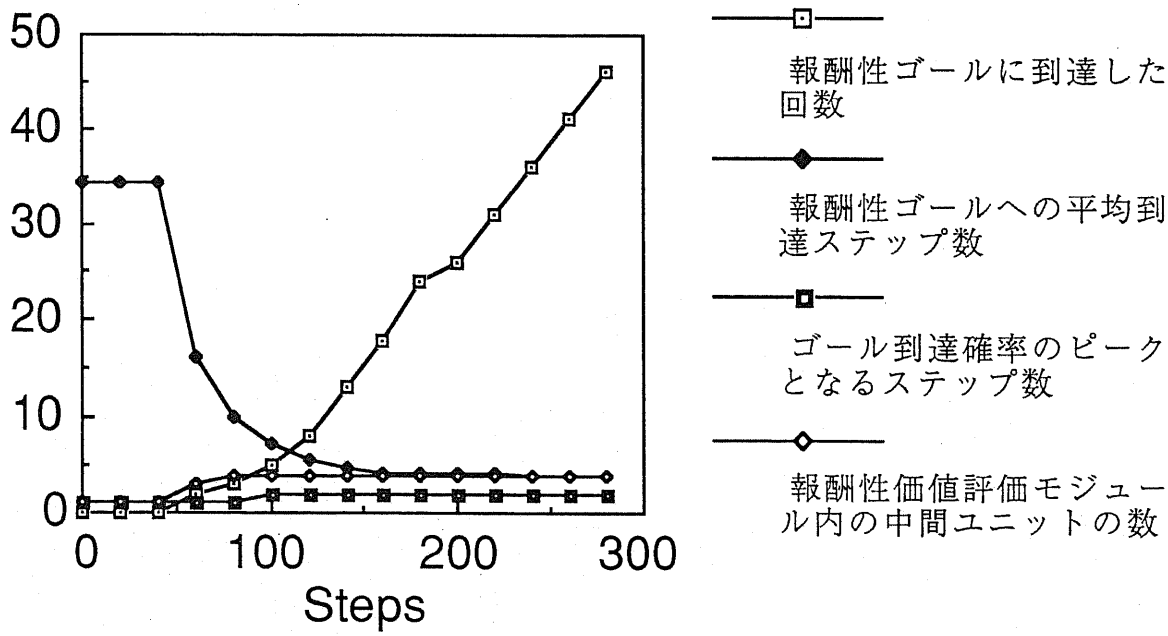
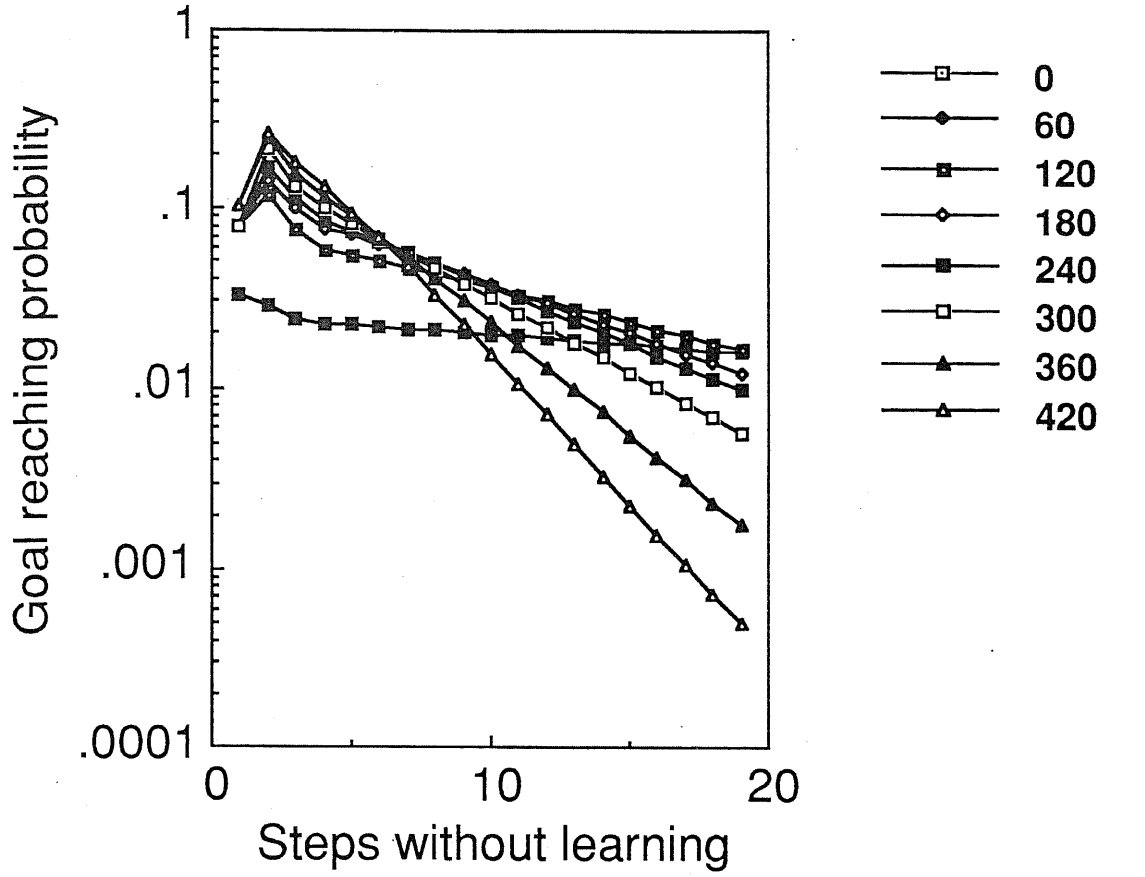


Fig. 4-15 一次モデルのシミュレーション (#1)

上図：学習を止めた状態で報酬性ゴールに到達する確率。
下図：学習に伴う動作主体の状態変化など。



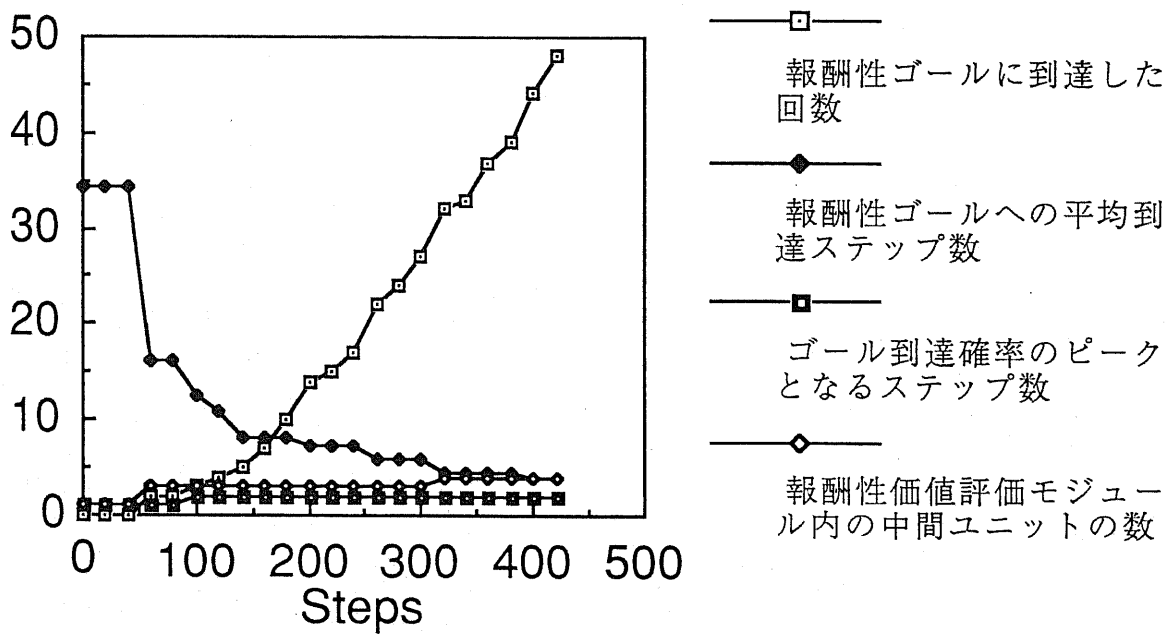


Fig. 4-16 一次モデルのシミュレーション (#2)

上図：学習を止めた状態で報酬性ゴールに到達する確率。
下図：学習に伴う動作主体の状態変化など。

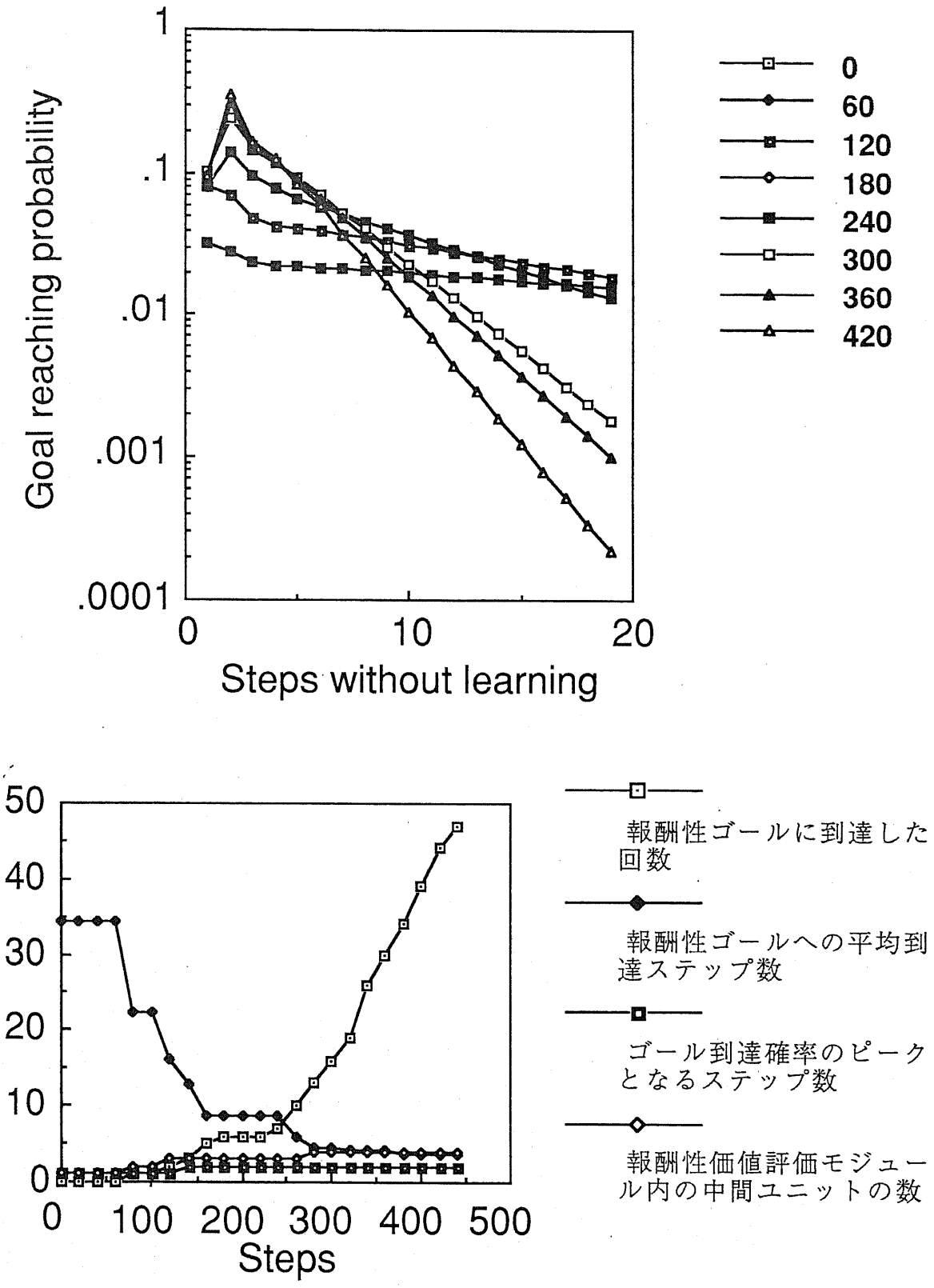
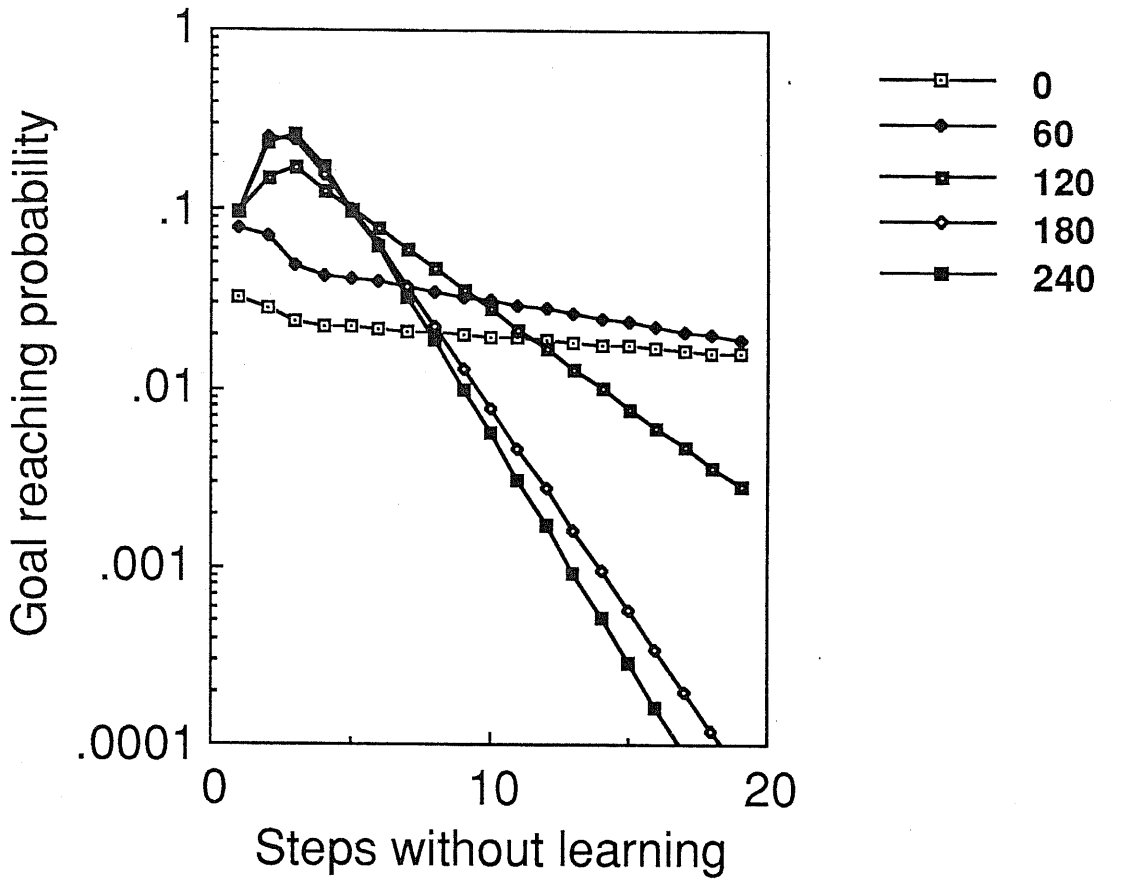


Fig. 4-17 一次モデルのシミュレーション (#3)

上図：学習を止めた状態で報酬性ゴールに到達する確率。
下図：学習に伴う動作主体の状態変化など。



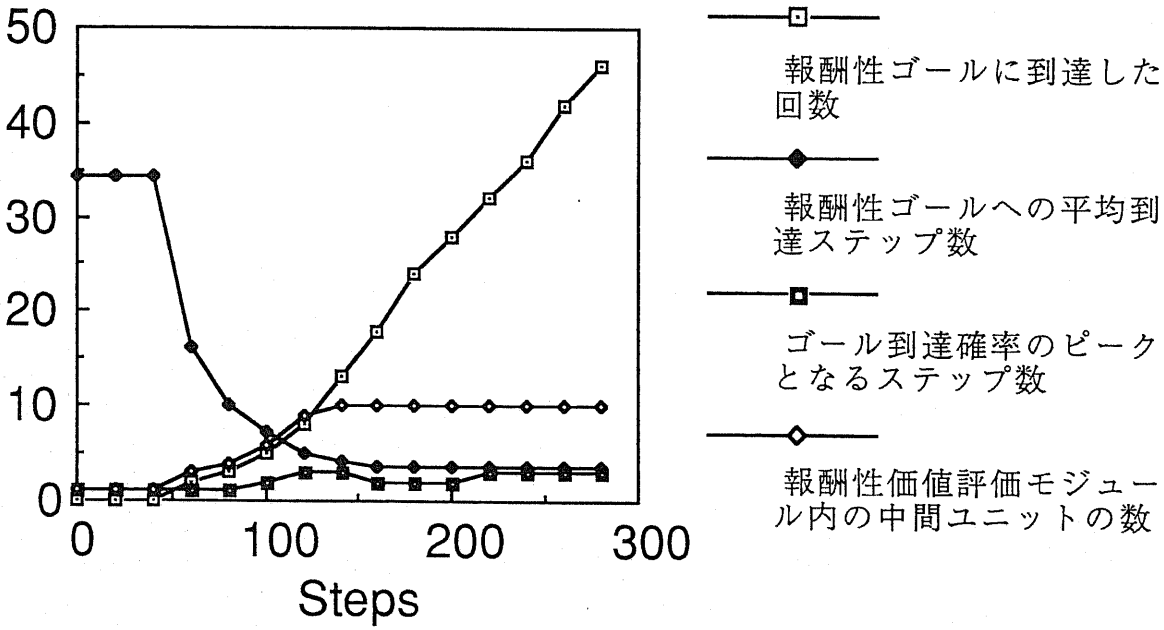
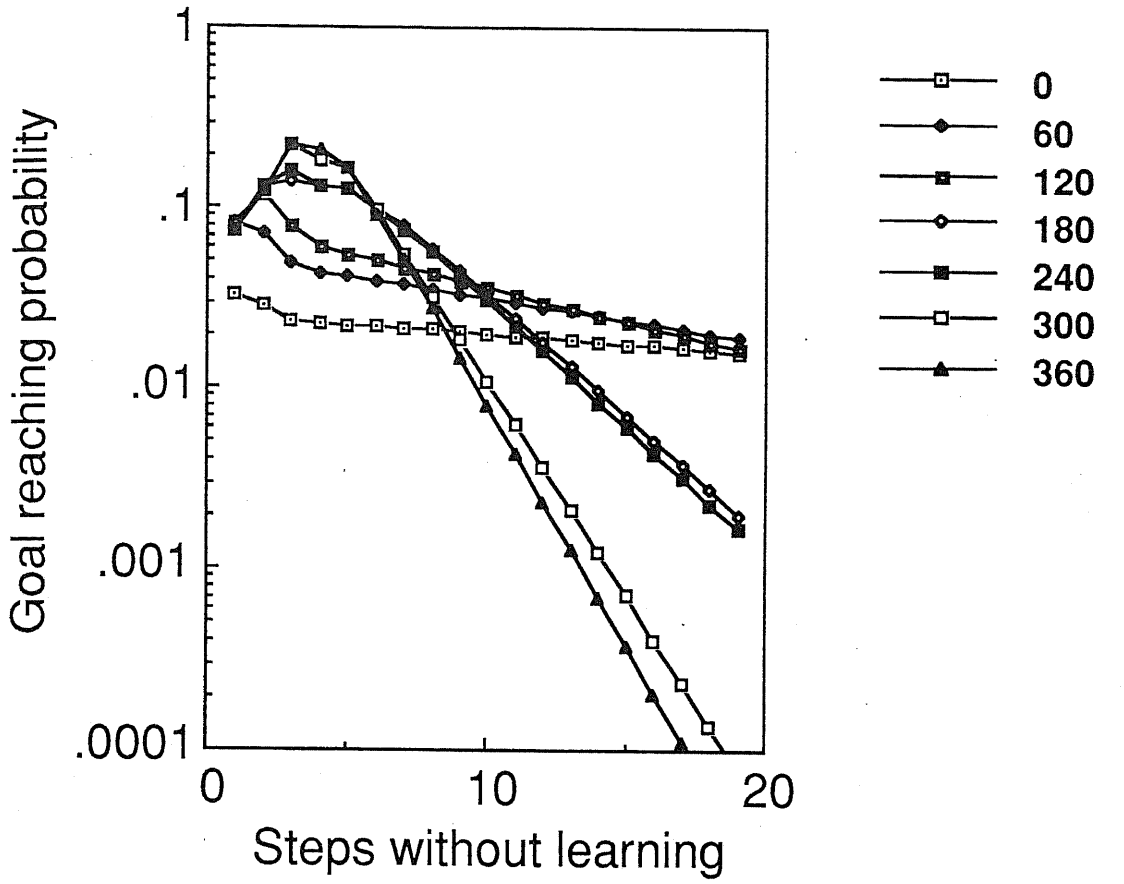


Fig. 4-18 再帰モデルのシミュレーション (#1)

上図：学習を止めた状態で報酬性ゴールに到達する確率。
下図：学習に伴う動作主体の状態変化など。



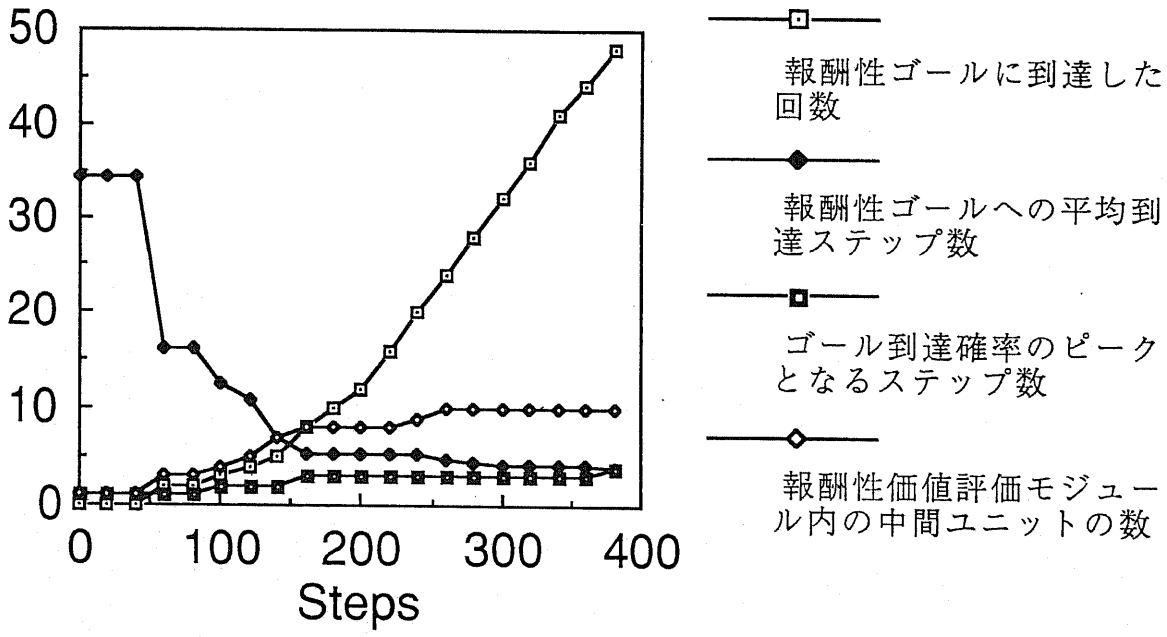
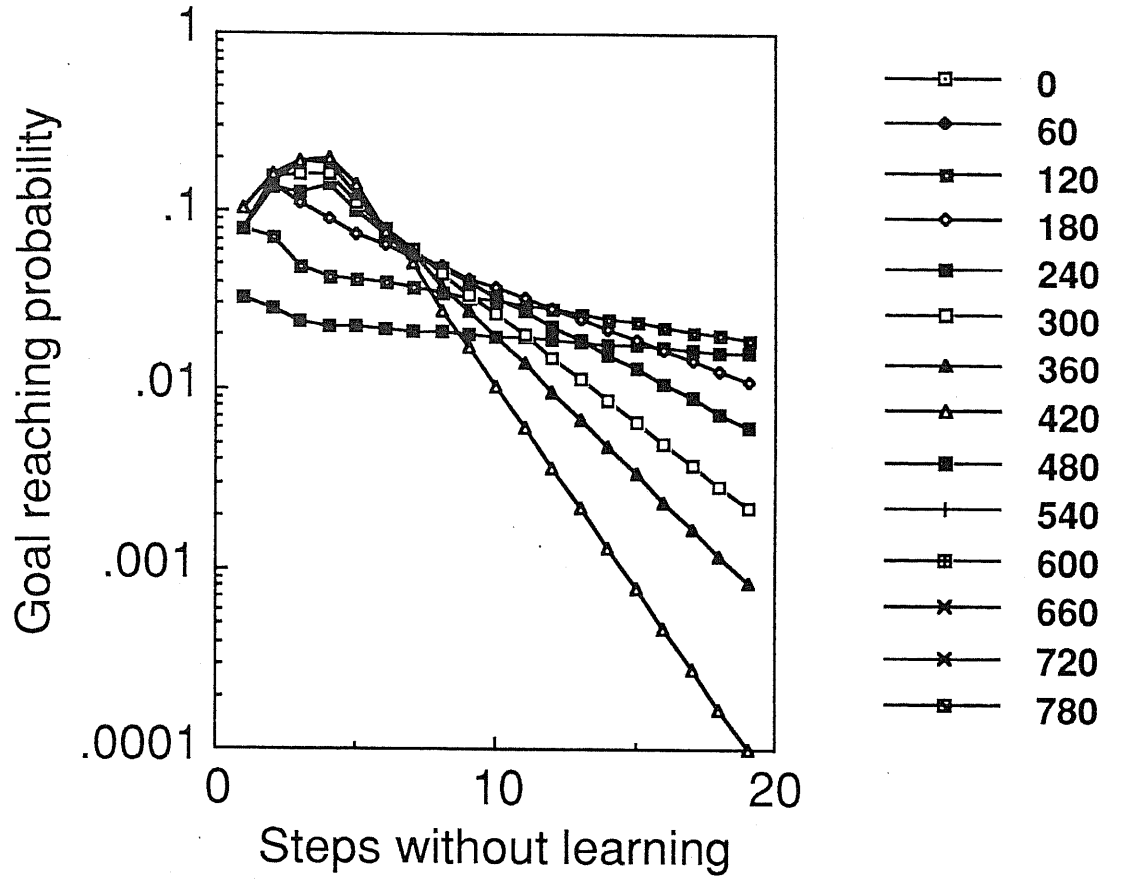


Fig. 4-19 再帰モデルのシミュレーション (#2)

上図：学習を止めた状態で報酬性ゴールに到達する確率。
下図：学習に伴う動作主体の状態変化など。



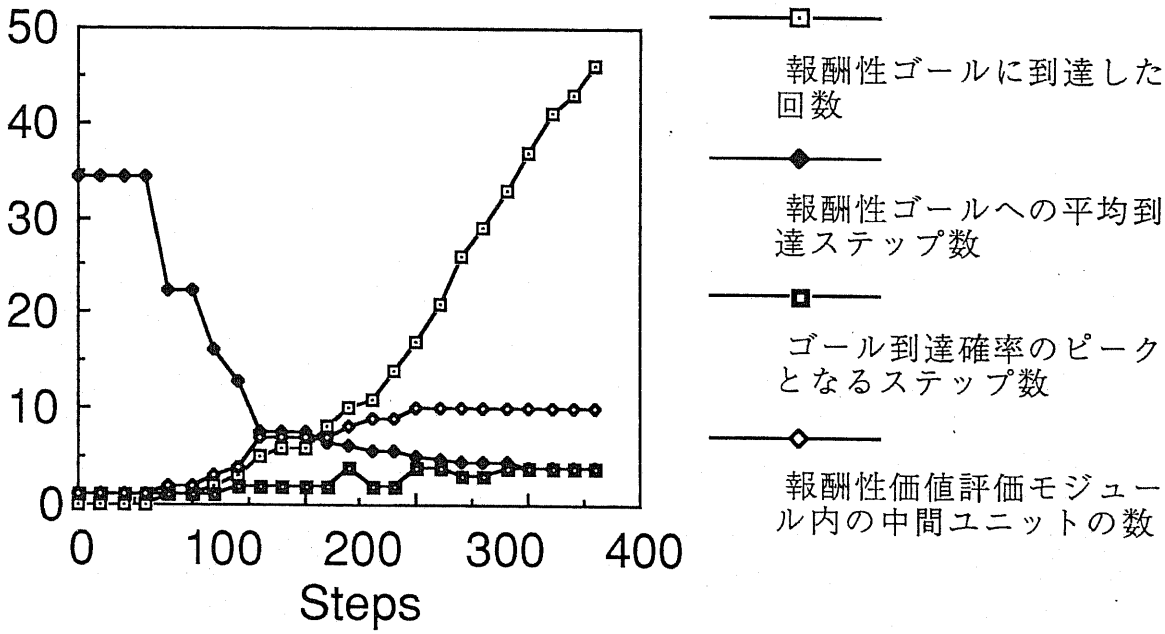
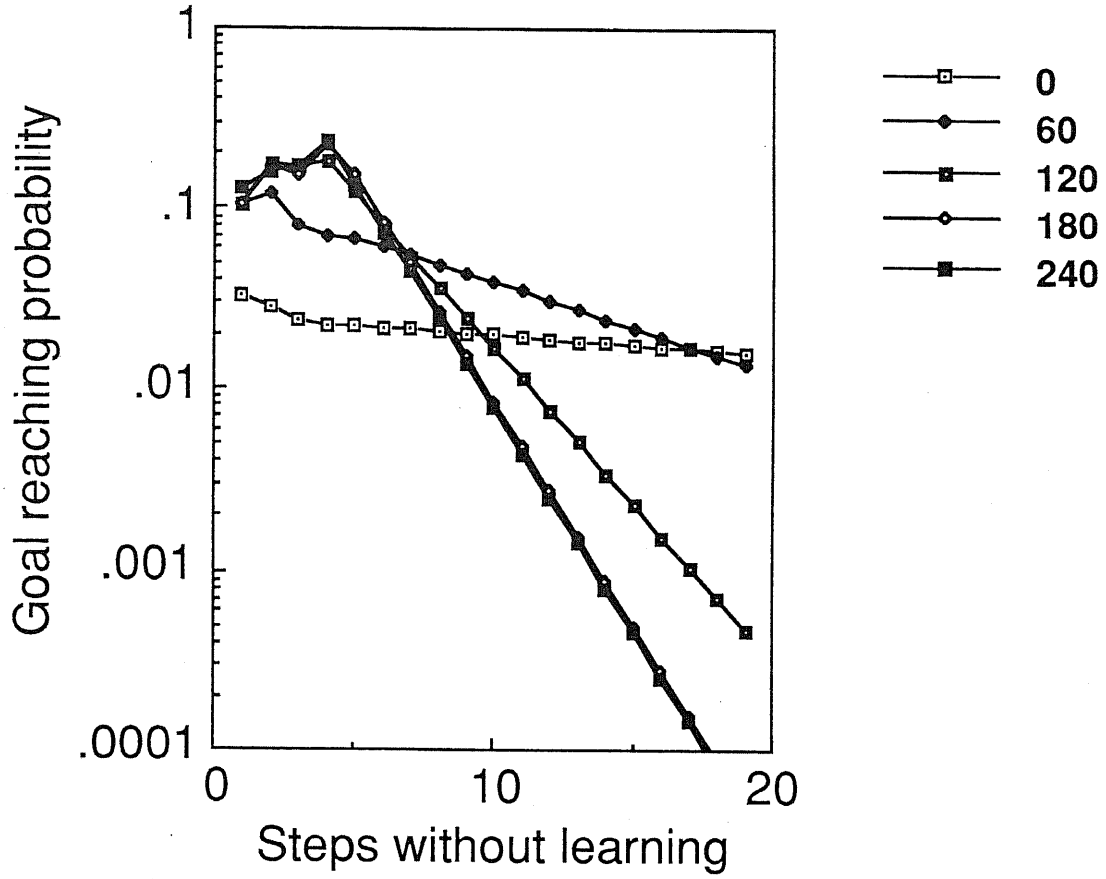


Fig. 4-20 再帰モデルのシミュレーション (#3)

上図：学習を止めた状態で報酬性ゴールに到達する確率。
下図：学習に伴う動作主体の状態変化など。



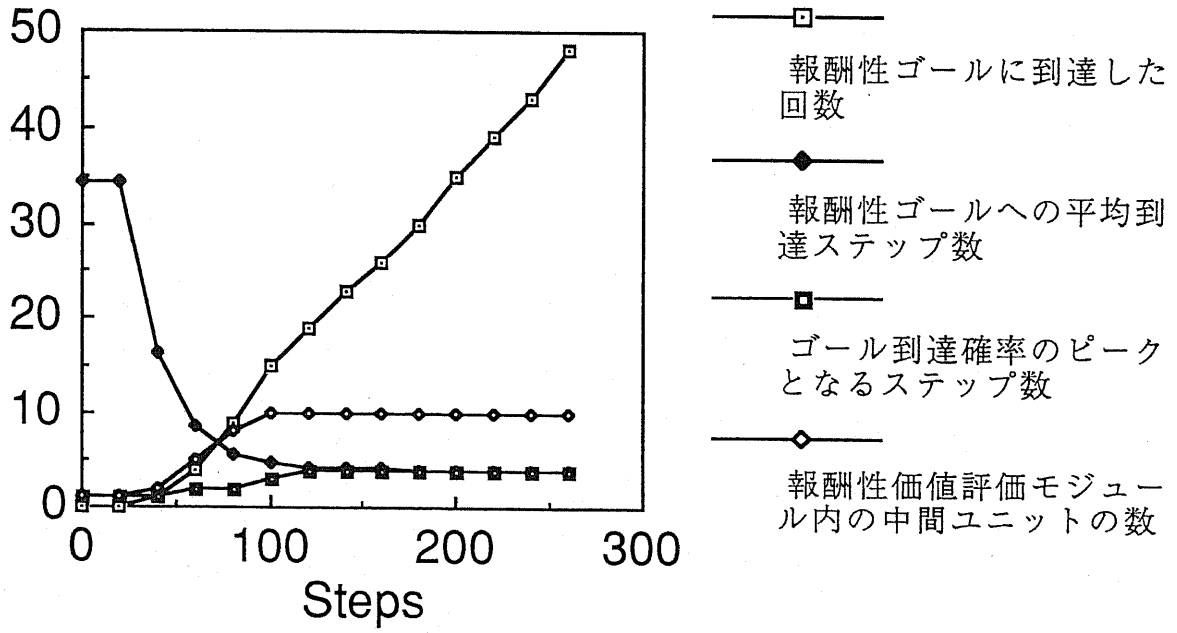
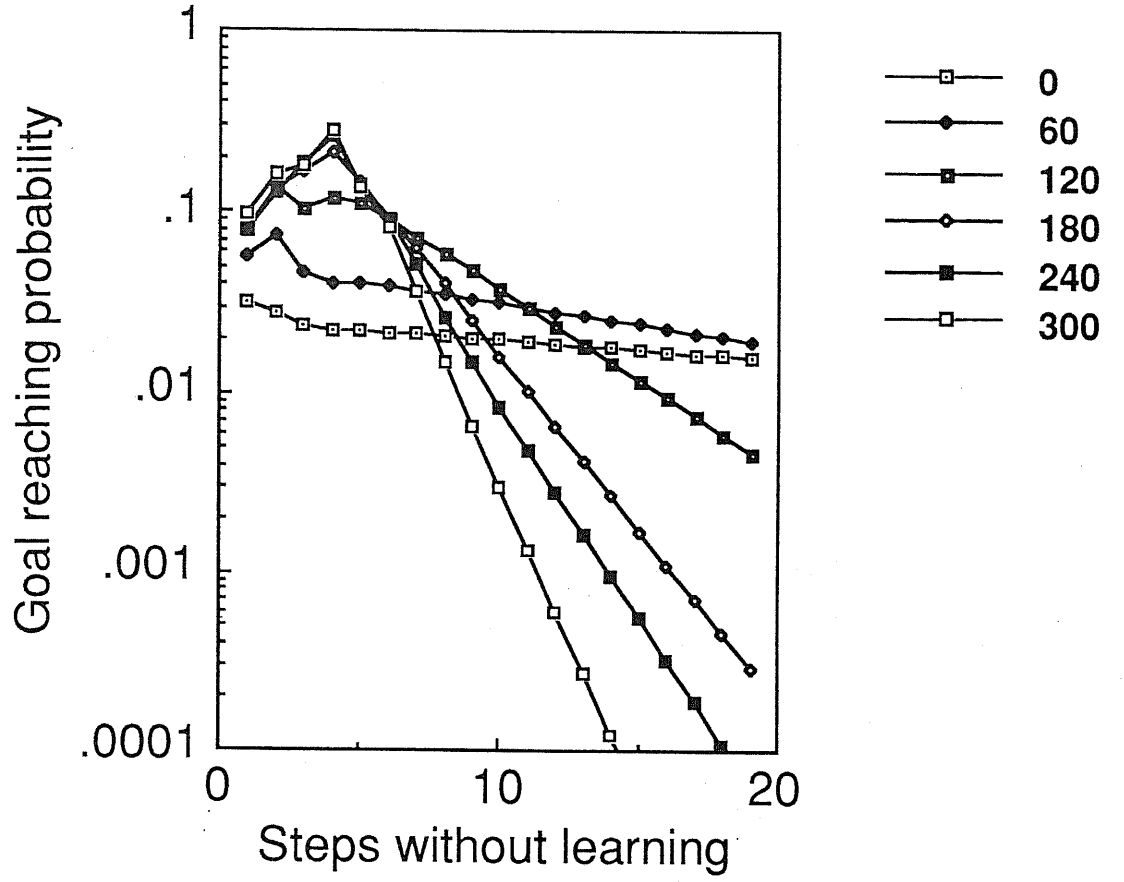


Fig. 4-21 再帰モデルのシミュレーション (#4)

上図：学習を止めた状態で報酬性ゴールに到達する確率。
下図：学習に伴う動作主体の状態変化など。



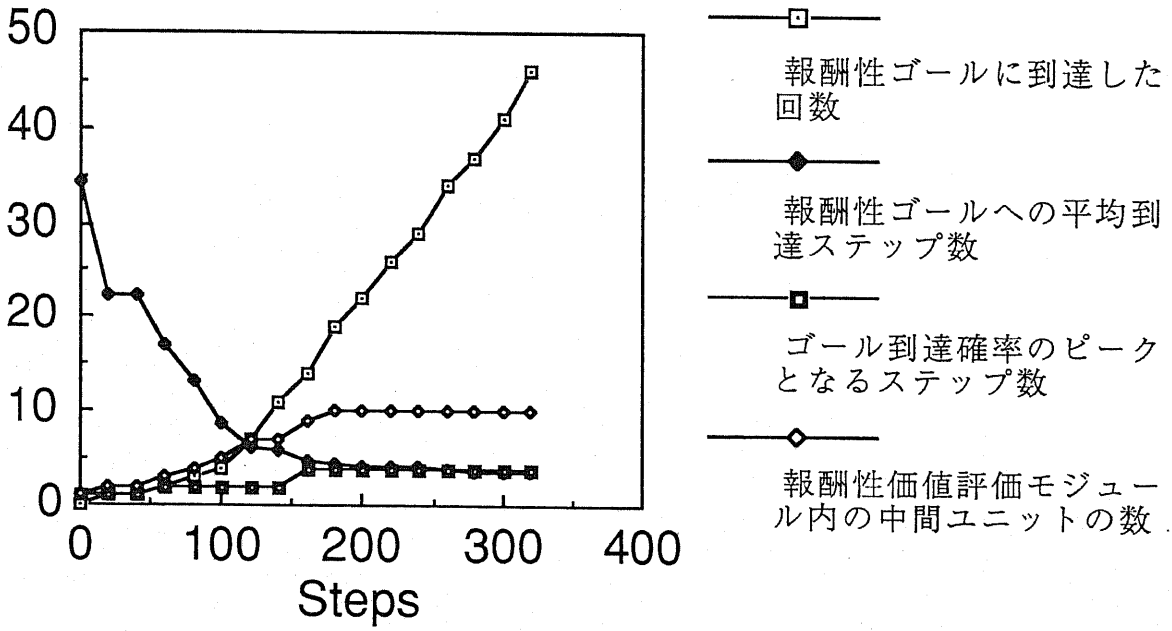


Fig. 4-22 再帰モデルのシミュレーション (#5)

上図：学習を止めた状態で報酬性ゴールに到達する確率。
下図：学習に伴う動作主体の状態変化など。

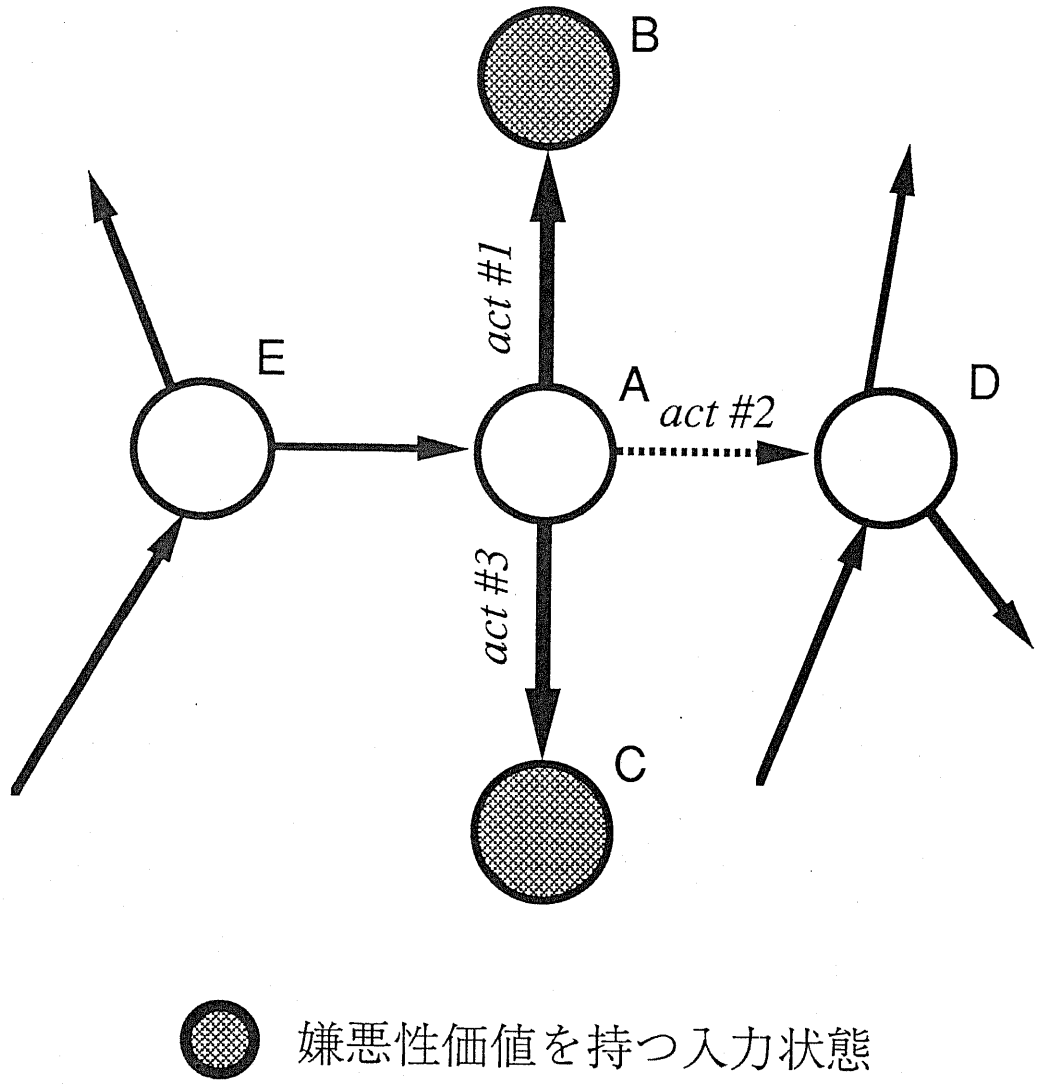


Fig. 4-23 逆固定型学習の問題点

図は動作主体と相互作用する環境を記述し、○は入力状態で、矢印は行動に毎に定義される入力状態の遷移ベクトルである。行動１(act#1)と行動３(act#3)がパターン的に逆である。

逆固定型の学習では、あるとき状態Ａに至り行動１を行うと、次の機会に再び状態Ａに至ると行動１を避けようとして次に状態Ａに到着したときには行動３を選ぶ。しかし行動３によっても嫌悪性が増大したので再び行動１を選ぶ。このようにして交互に行動を行うだけで適切な回避行動２を選ぶことができない。

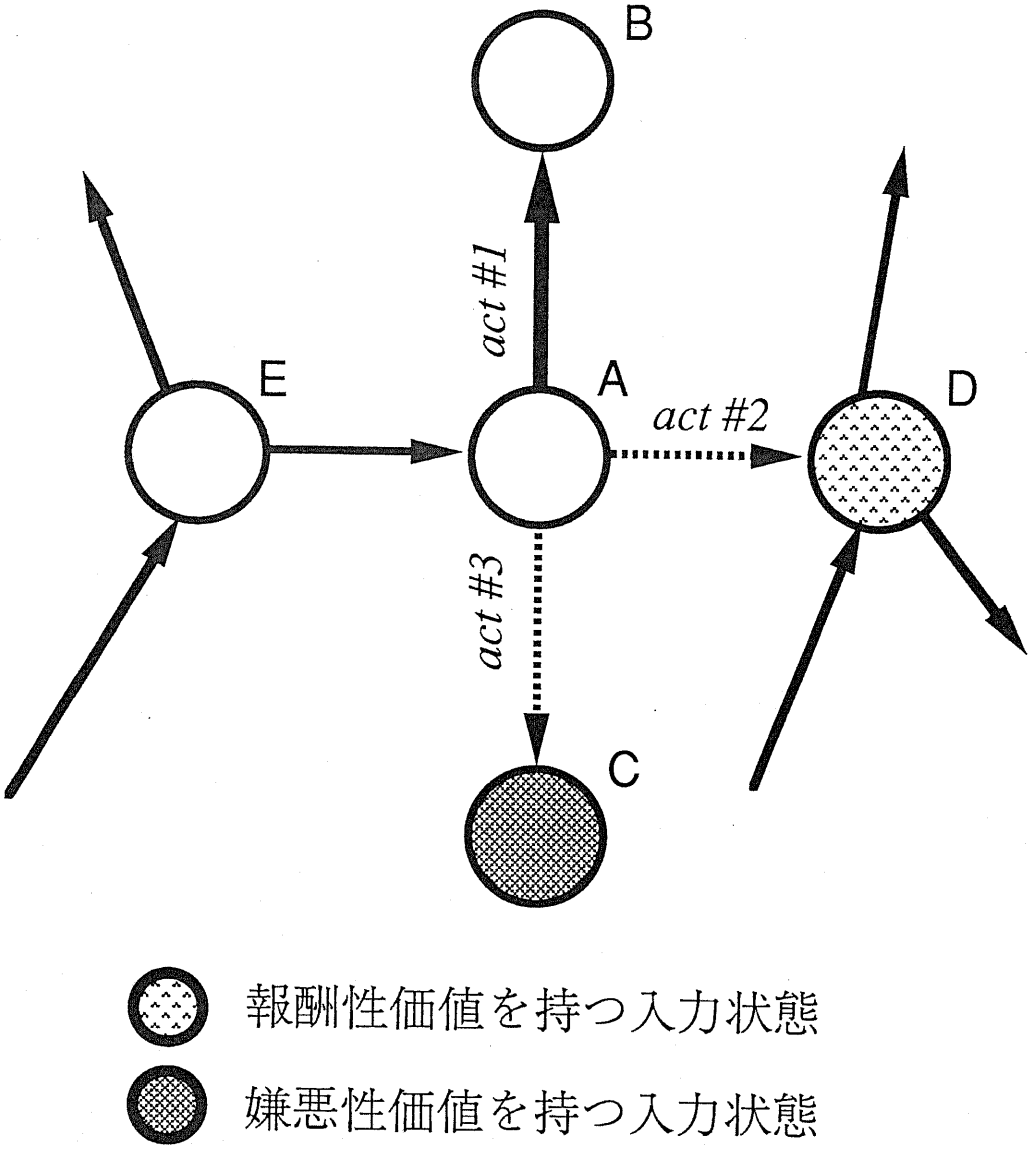
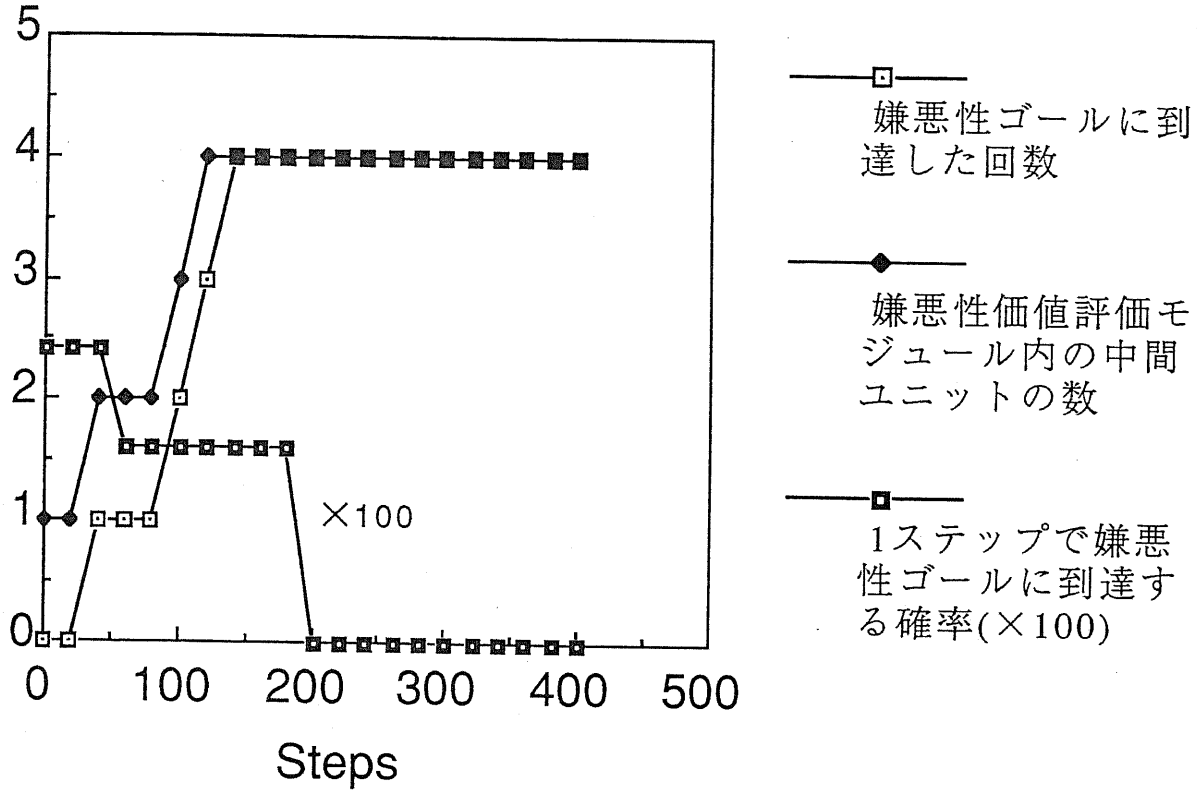


Fig. 4-24 逆固定型学習の問題点

図は動作主体と相互作用する環境を記述し、○は入力状態で、矢印は行動に毎に定義される入力状態の遷移ベクトルである。行動１(act#1)と行動３(act#3)がパターン的に逆である。

逆固定型学習では嫌悪性の状態Ｃを避けようとしてパターン的に逆の行動１を選んでしまうと、より優れた行動２を選ぶことが出来ない。脱出記憶型では行動パターンには関係無いが同様の問題が発生し得る。



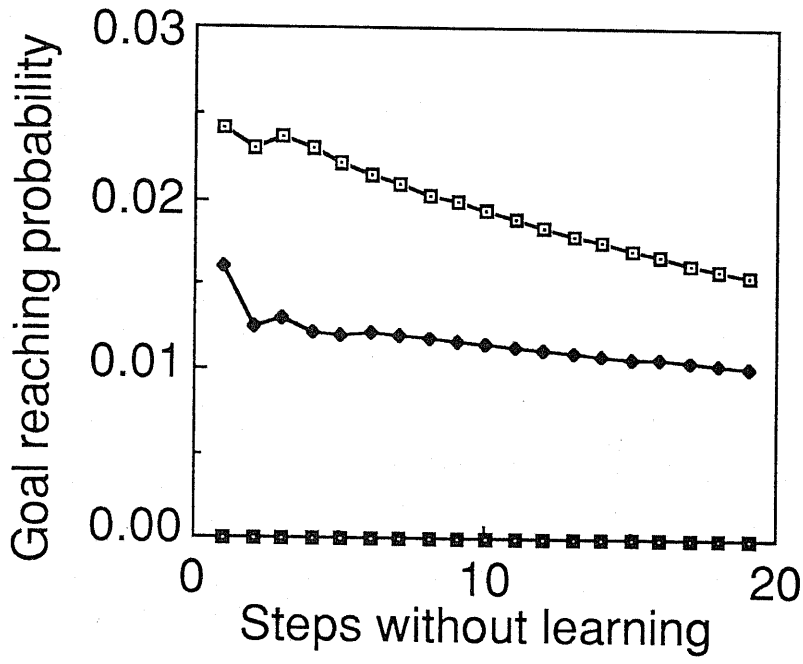
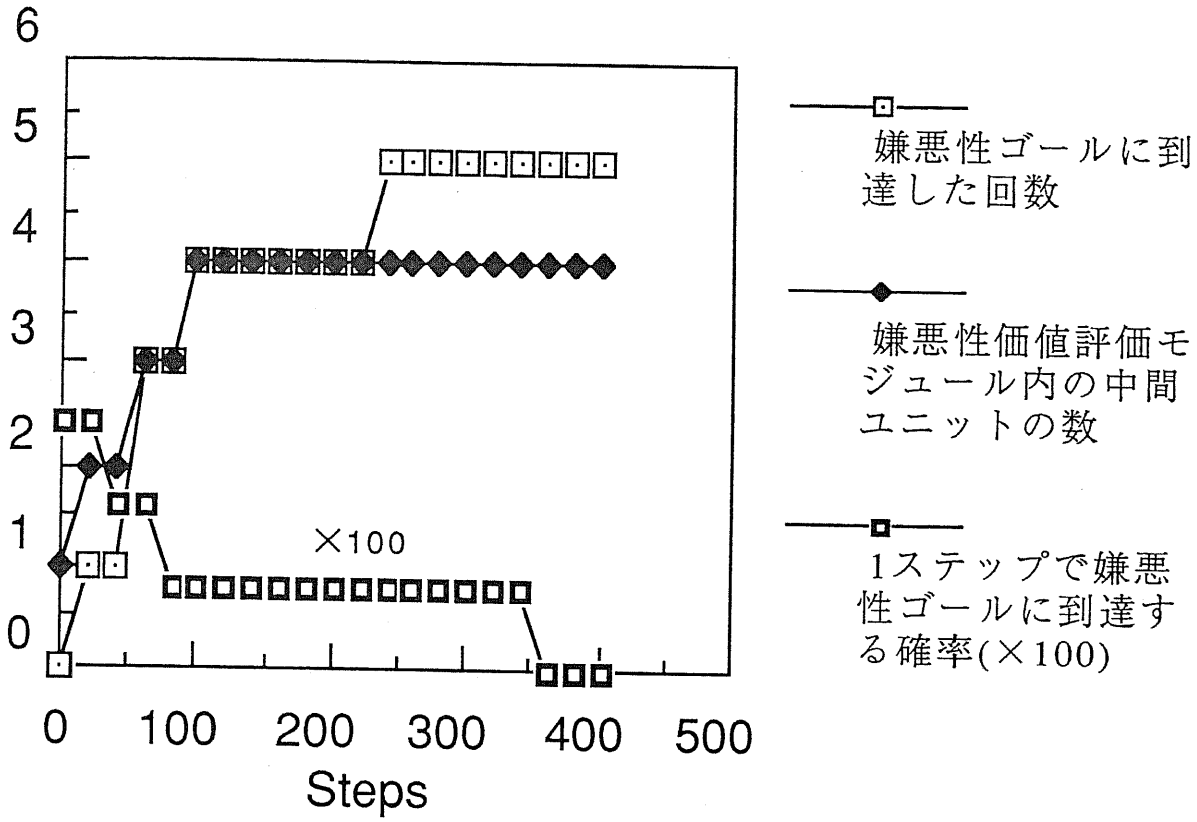


Fig. 4-25嫌悪性のみによるシミュレーション(#1)

上図：学習に伴う動作主体の状態変化など。下図：学習を止めた状態で嫌悪性ゴールに到達する確率。上図の１ステップでゴールに到達する確率が減少すると下図の曲線が変化する。



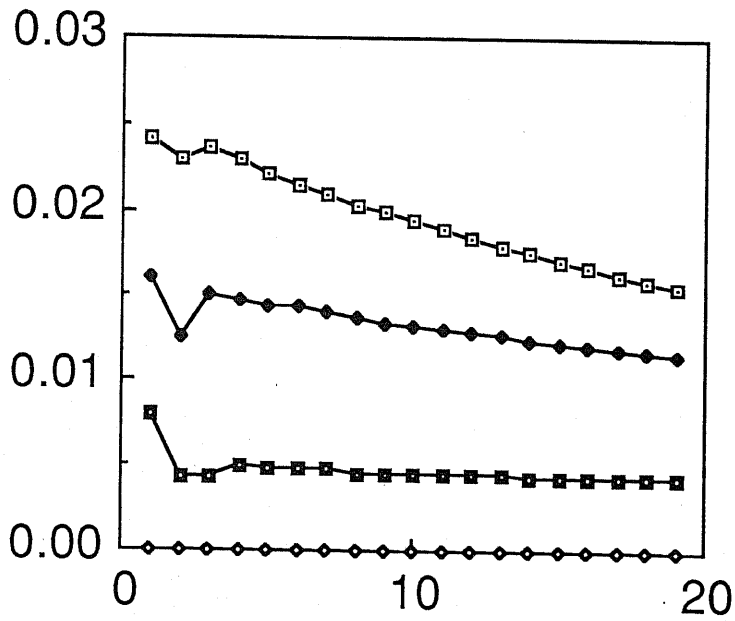
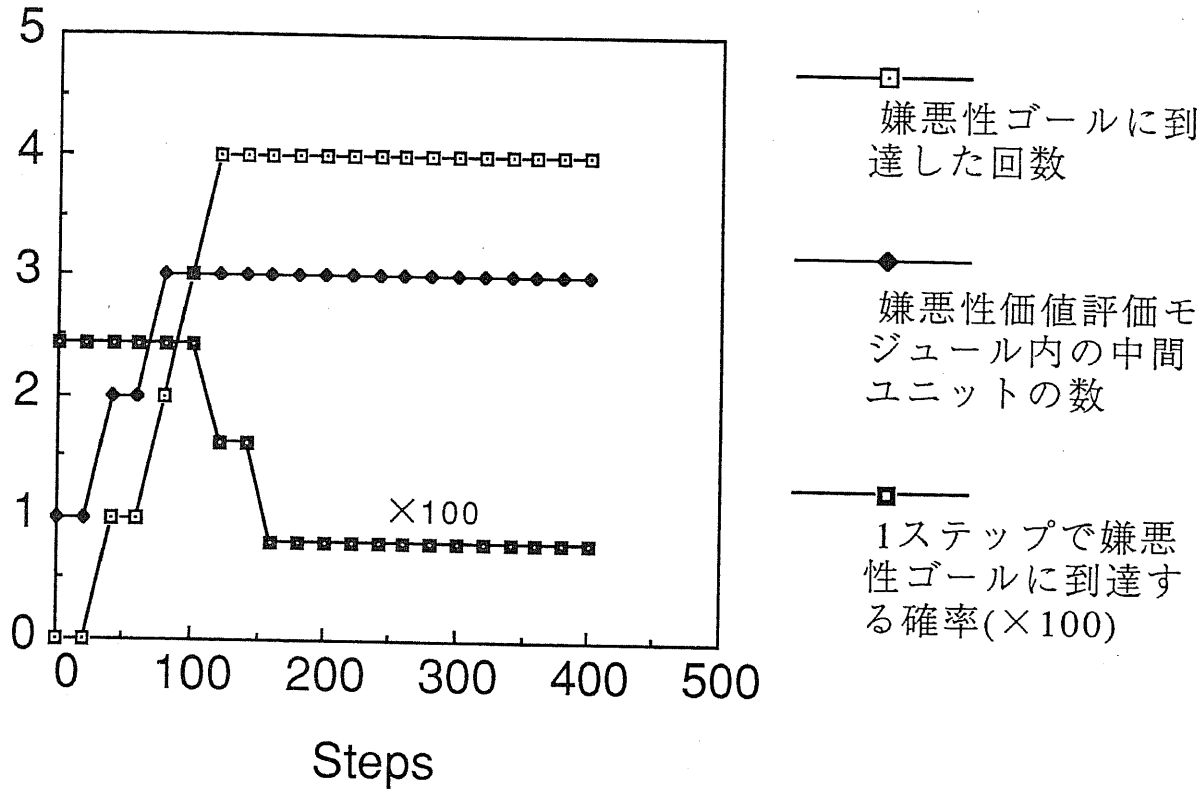


Fig. 4-26 嫌悪性のみによるシミュレーション(#2)

上図：学習に伴う動作主体の状態変化など。下図：学習を止めた状態で嫌悪性ゴールに到達する確率。上図の１ステップでゴールに到達する確率が減少すると下図の曲線が変化する。



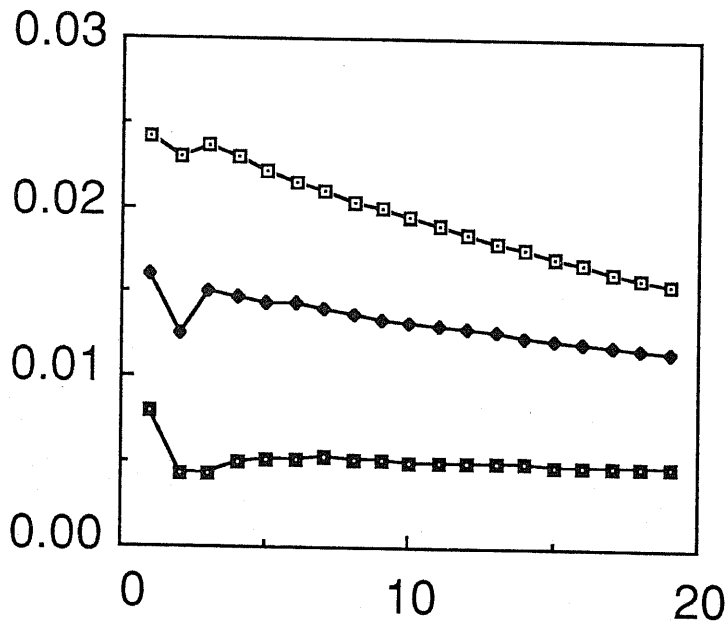
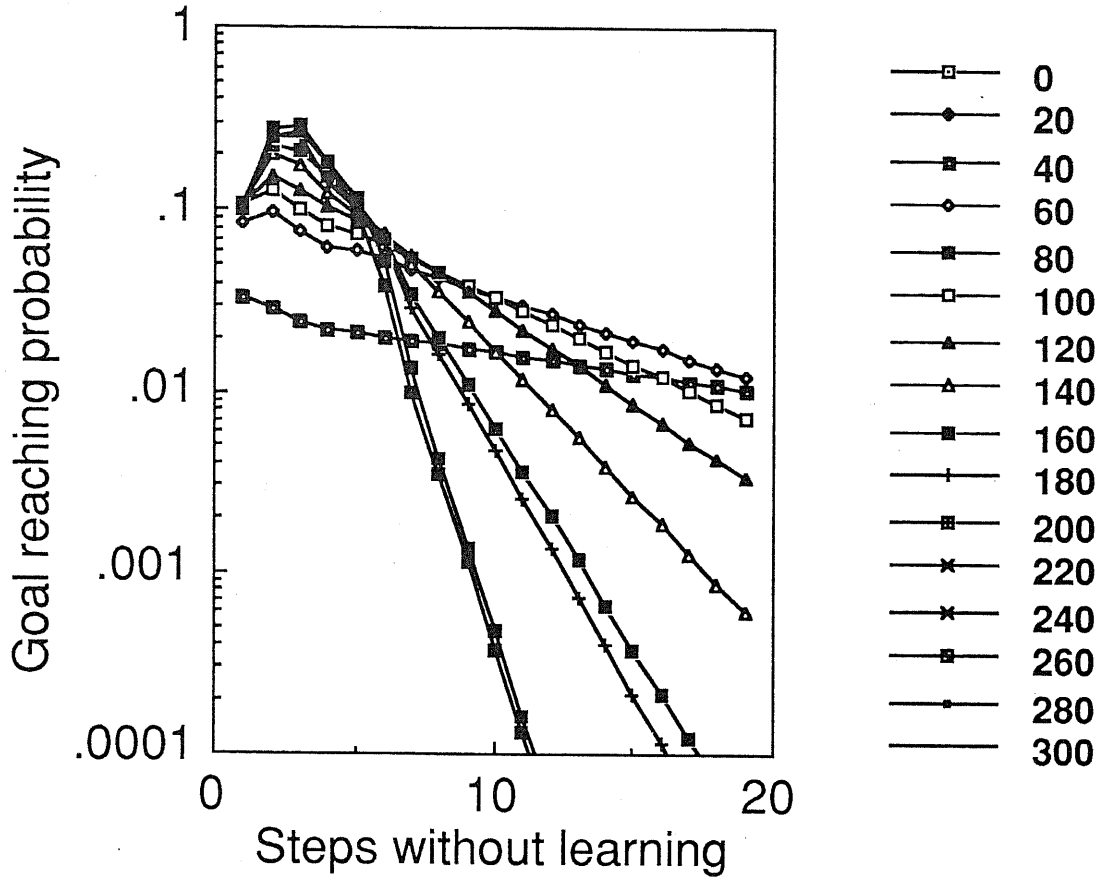


Fig. 4-27嫌悪性のみによるシミュレーション(#3)

上図：学習に伴う動作主体の状態変化など。下図：学習を止めた状態で嫌悪性ゴールに到達する確率。上図の１ステップでゴールに到達する確率が減少すると下図の曲線が変化する。



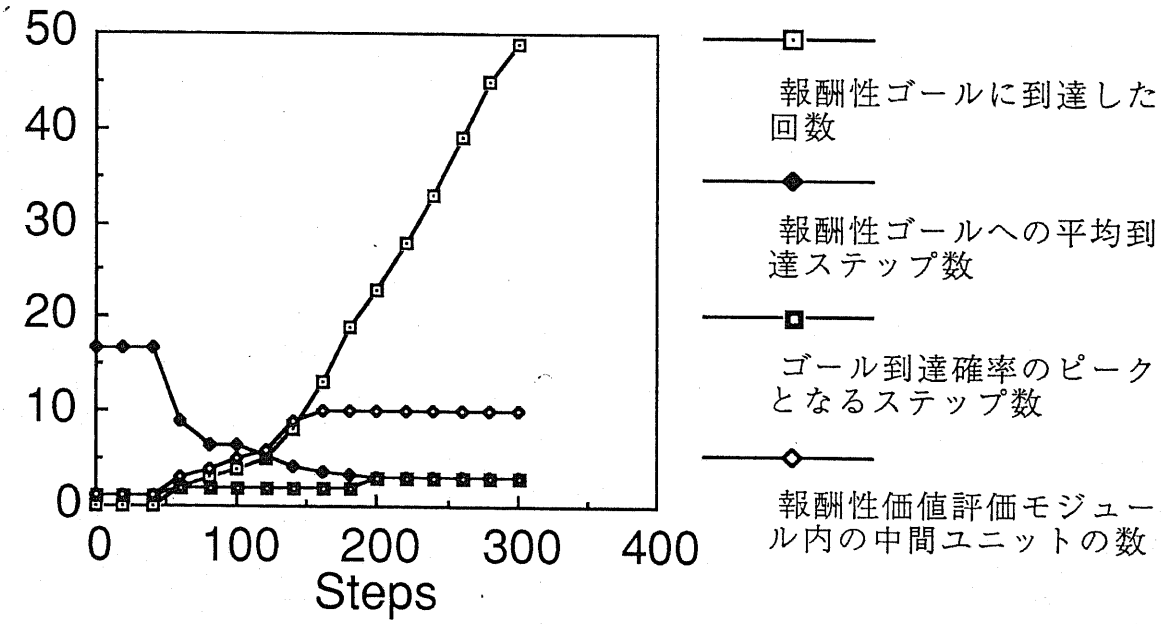
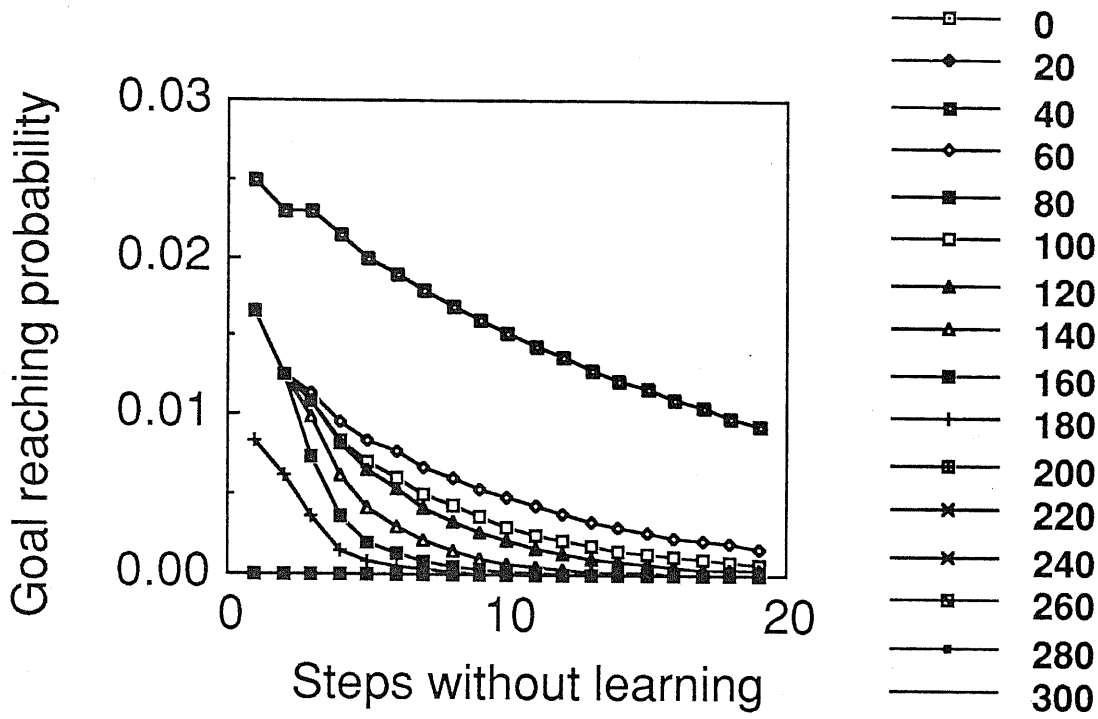


Fig. 4-28(a) 報酬性と嫌悪性によるシミュレーション (#1)
[報酬性ゴールに対する評価]

上図：学習を止めた状態で報酬性ゴールに到達する確率。
下図：学習に伴う動作主体の状態変化など。



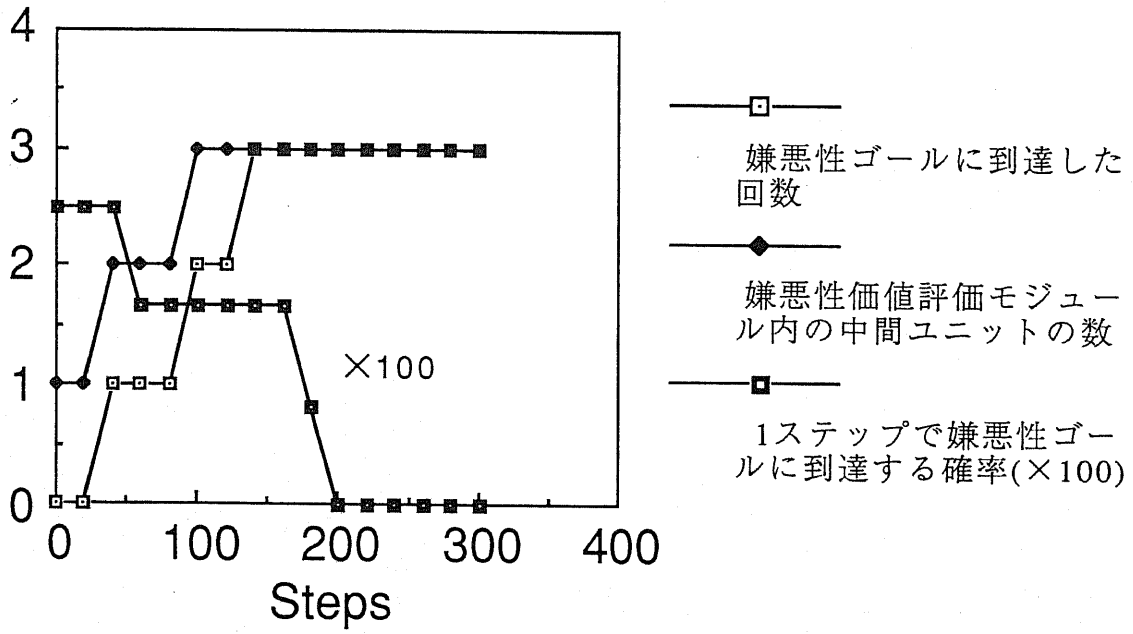
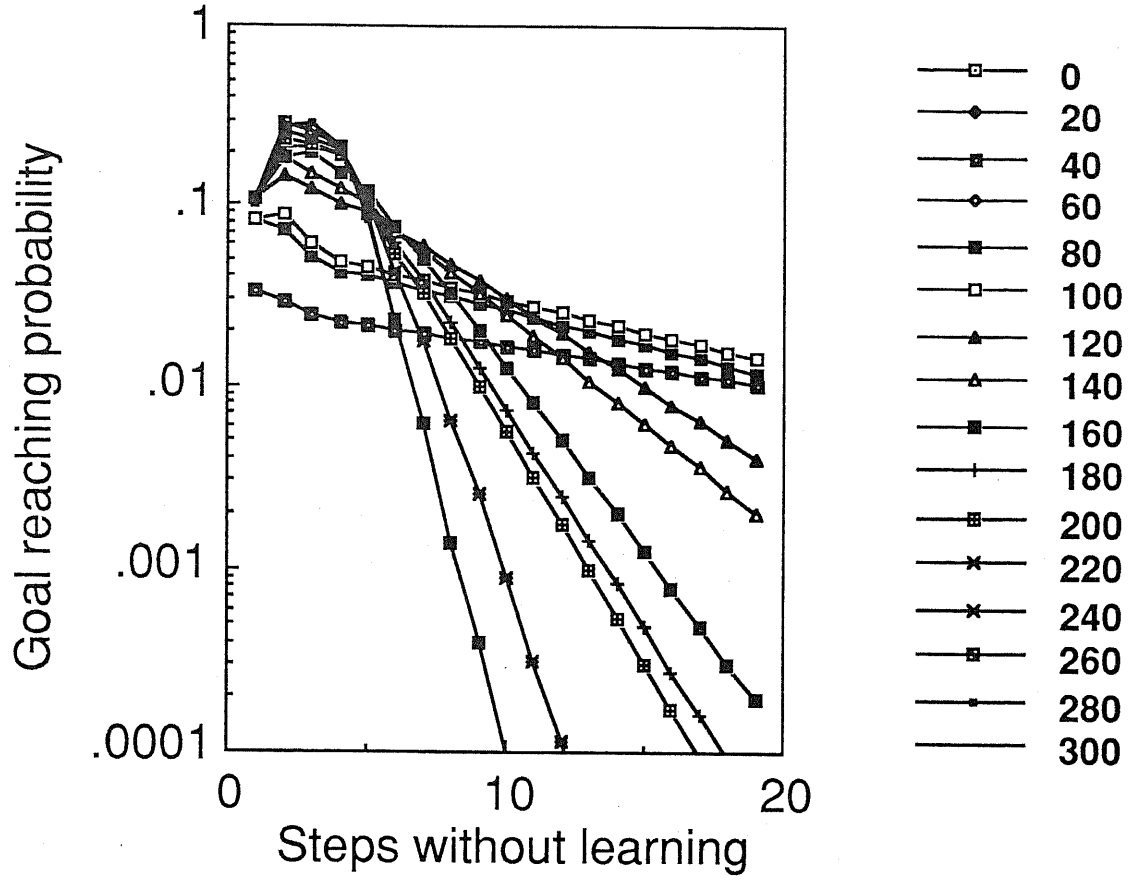


Fig. 4-28(b) 報酬性と嫌悪性によるシミュレーション (#1)
[嫌悪性ゴールに対する評価]

上図：学習を止めた状態で嫌悪性ゴールに到達する確率。
下図：学習に伴う動作主体の状態変化など。



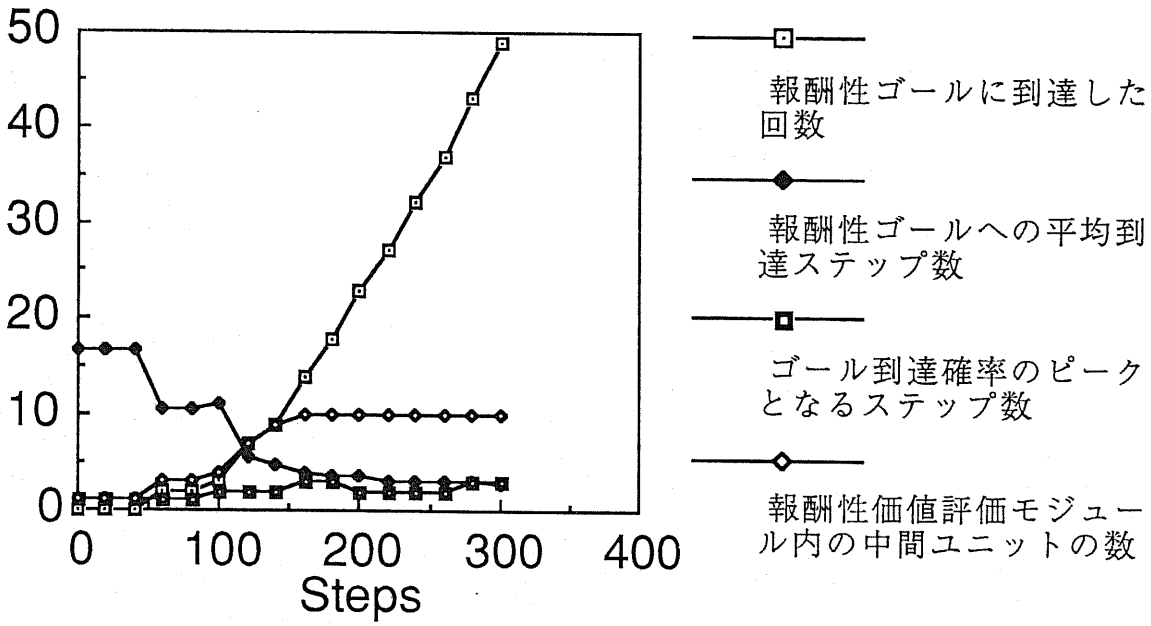
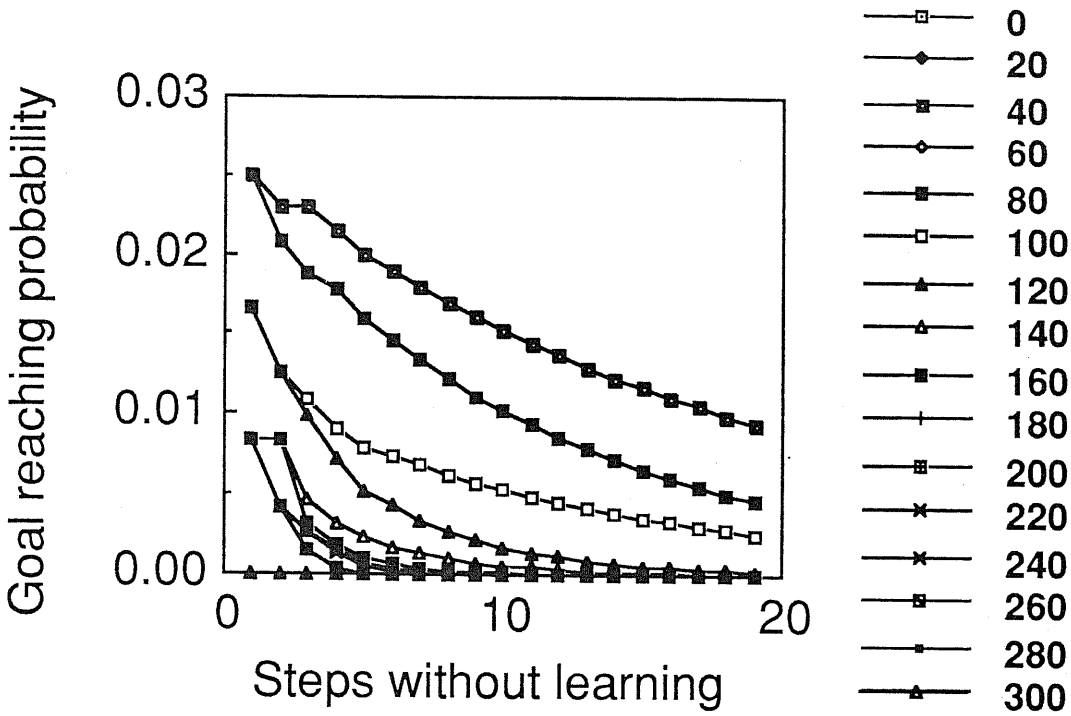


Fig. 4-29 (a) 報酬性と嫌悪性によるシミュレーション (#2)
[報酬性ゴールに対する評価]

上図：学習を止めた状態で報酬性ゴールに到達する確率。
下図：学習に伴う動作主体の状態変化など。



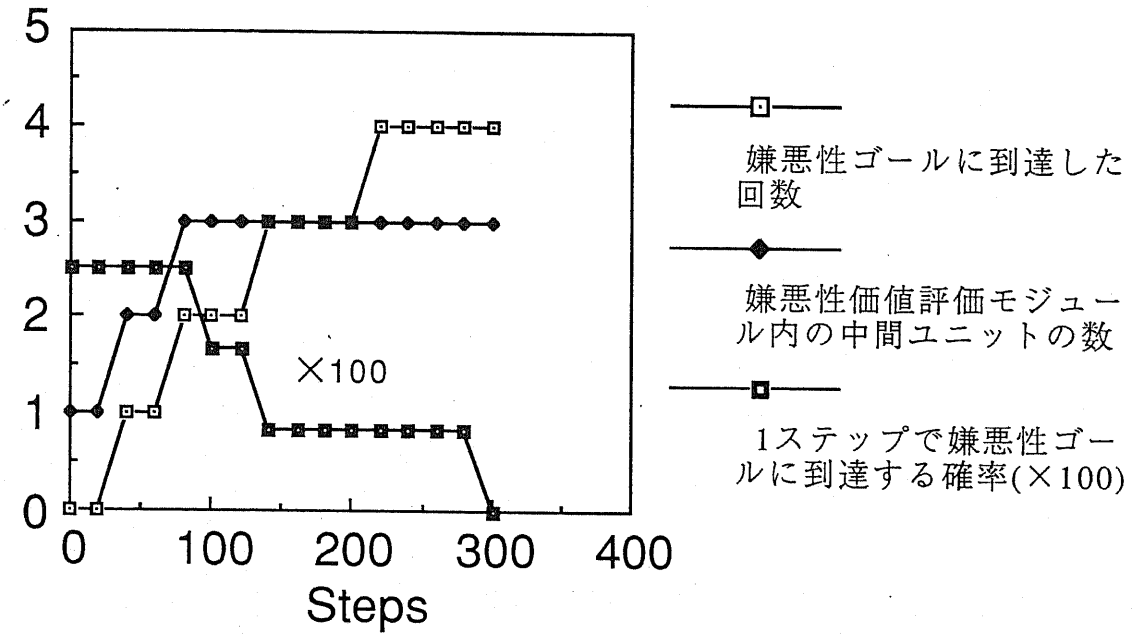
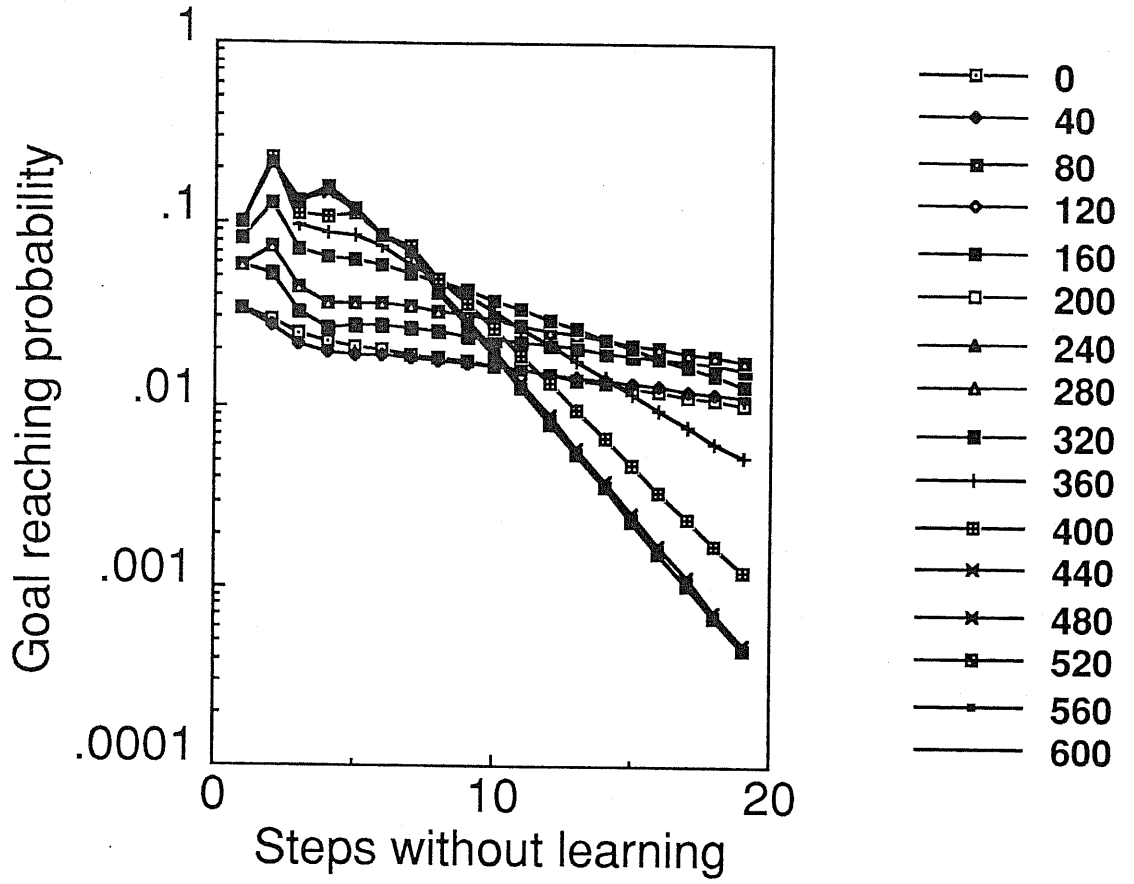


Fig. 4-29(b) 報酬性と嫌悪性によるシミュレーション (#2)
[嫌悪性ゴールに対する評価]

上図：学習を止めた状態で嫌悪性ゴールに到達する確率。
下図：学習に伴う動作主体の状態変化など。



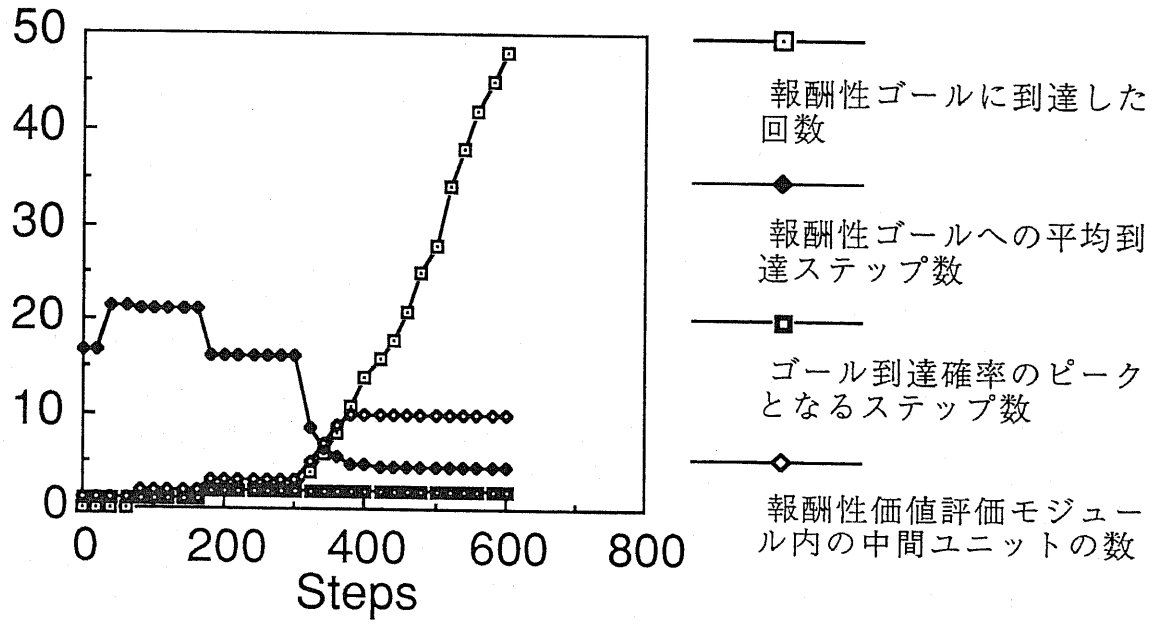
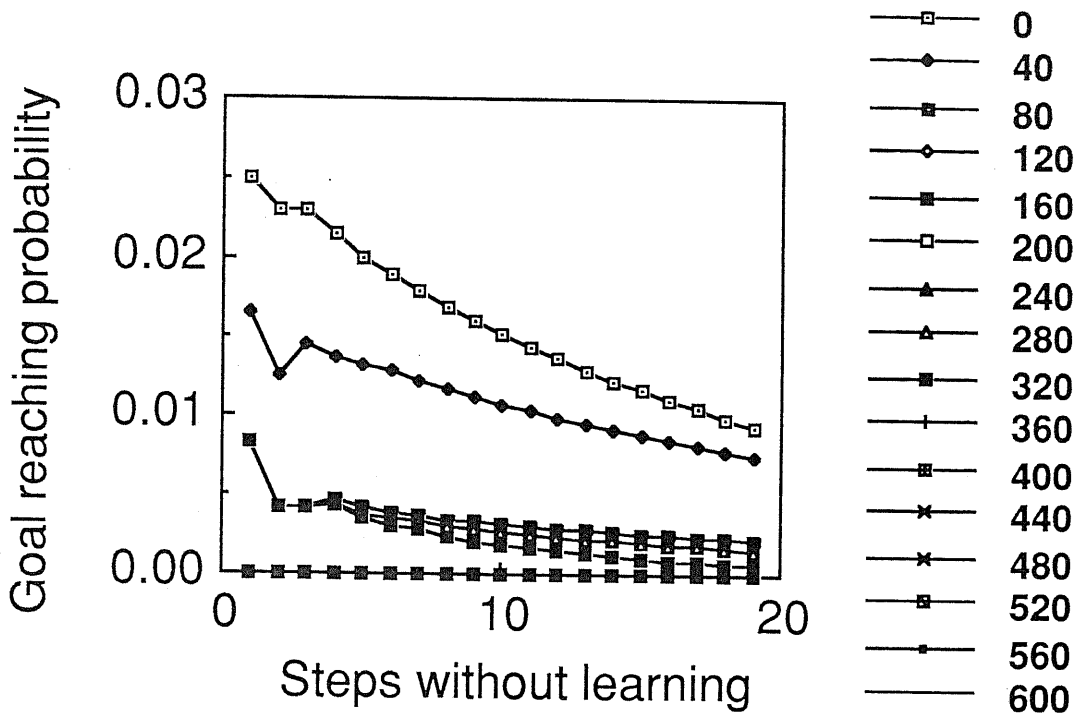


Fig. 4-30 (a) 報酬性と嫌悪性によるシミュレーション (#3)
[報酬性ゴールに対する評価]

上図：学習を止めた状態で報酬性ゴールに到達する確率。
下図：学習に伴う動作主体の状態変化など。



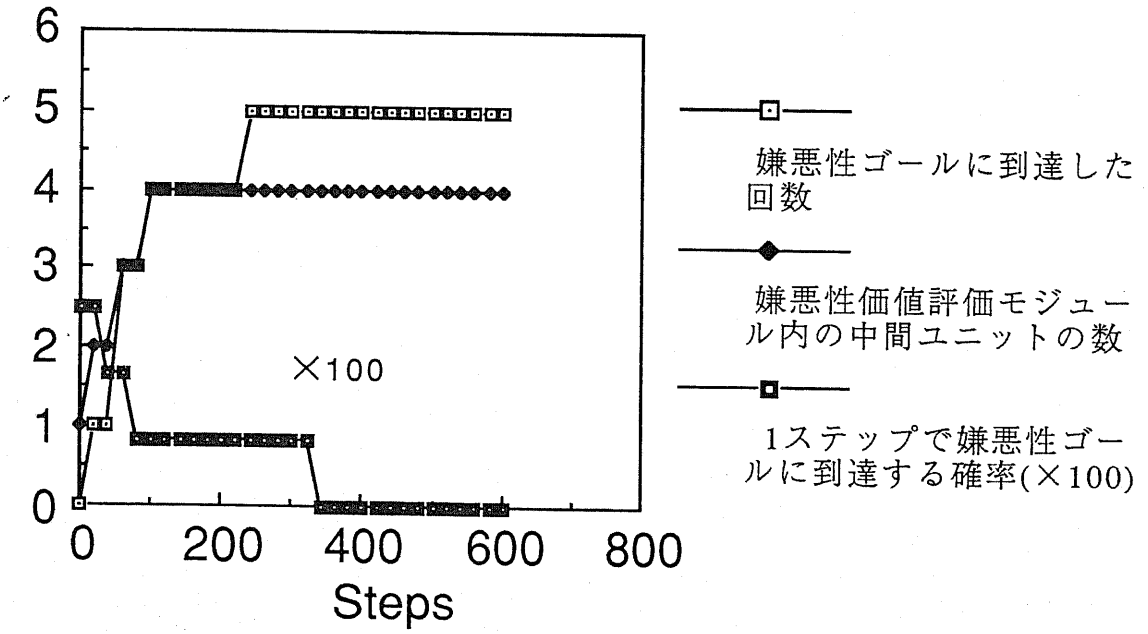
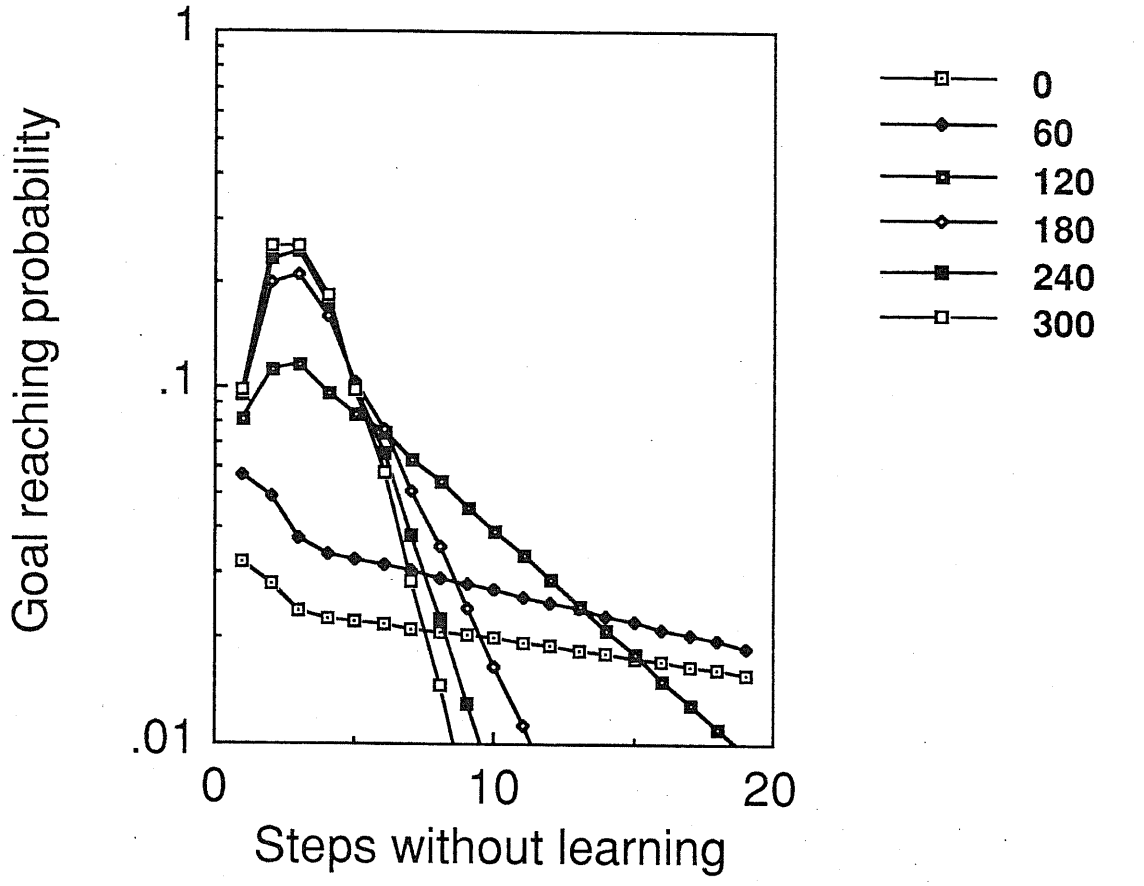


Fig. 4-30(b) 報酬性と嫌悪性によるシミュレーション (#3)
[嫌悪性ゴールに対する評価]

上図：学習を止めた状態で嫌悪性ゴールに到達する確率。
下図：学習に伴う動作主体の状態変化など。



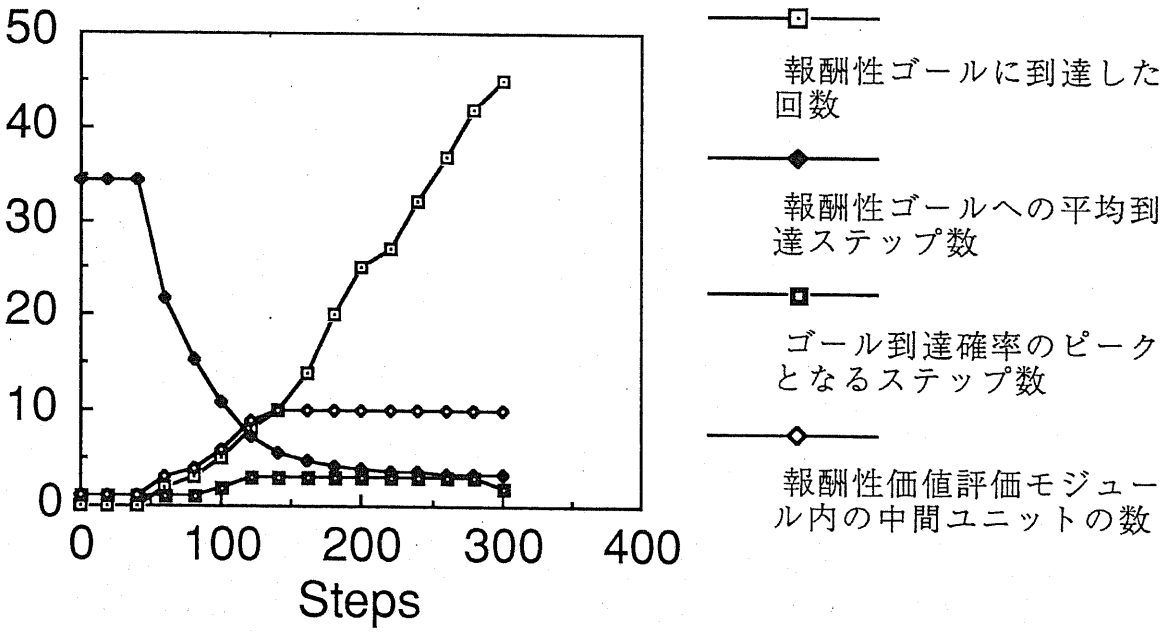


Fig. 4-3l(a) 確定的な環境における慎重な動作主体 (#1)
(報酬性ゴールのみを持つ環境)

上図：学習を止めた状態で報酬性ゴールに到達する確率。
下図：学習に伴う動作主体の状態変化など。

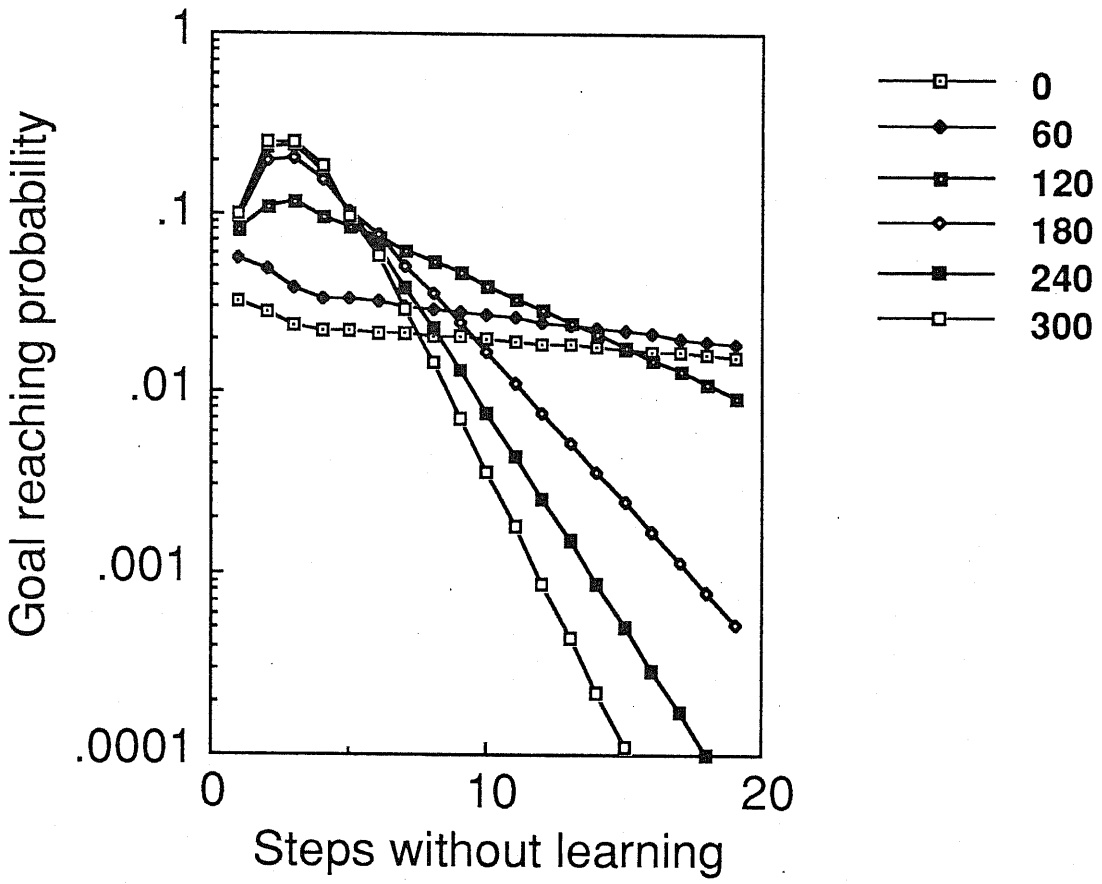
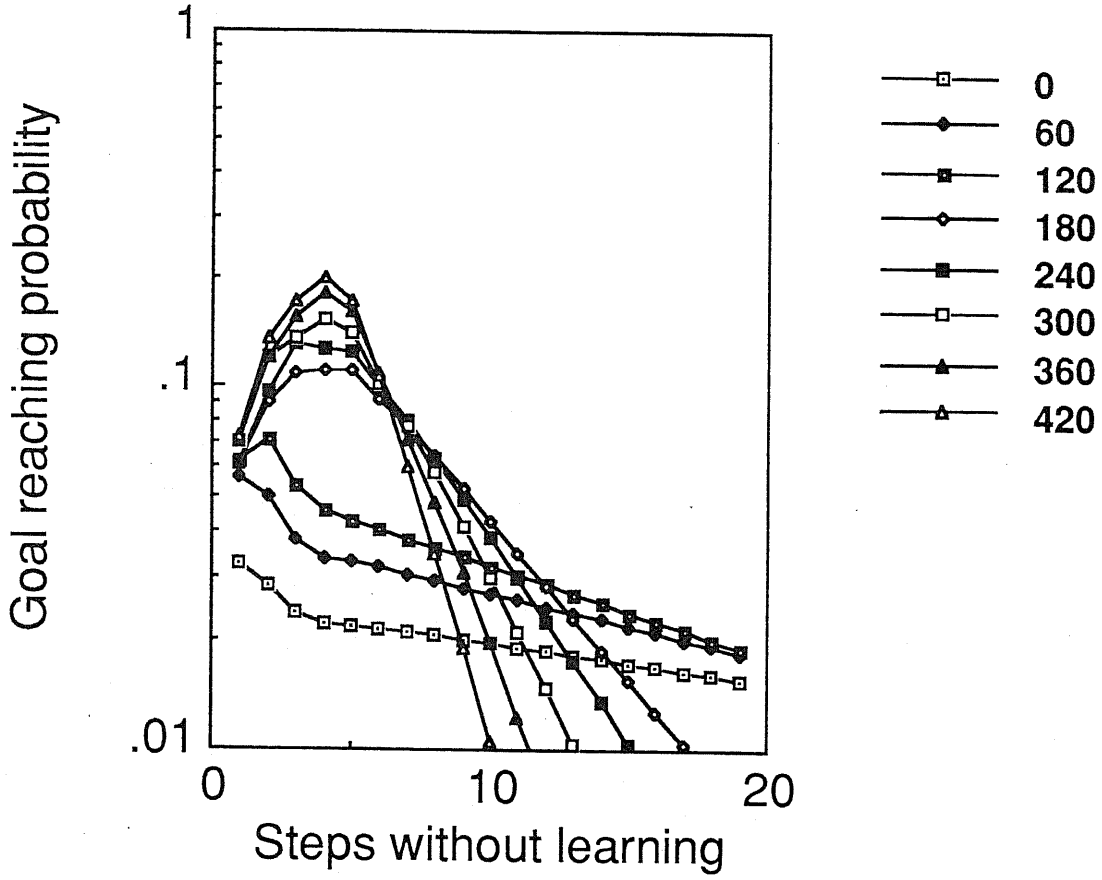


Fig. 4-3|(b) 確定的な環境における慎重な動作主体 (#1)
(報酬性ゴールのみを持つ環境)

この図はFig.4- (a)の上図と同じものだが、これまでの、シミュレーションの比較のために縦軸をlogスケールで0.0001から1.0までを表示した。



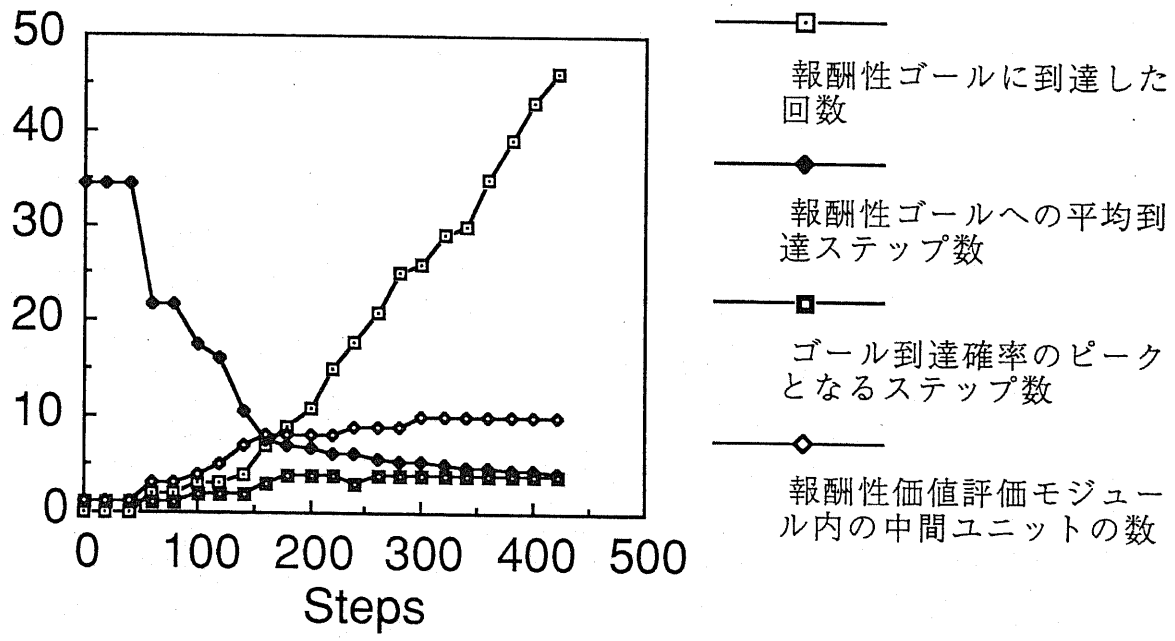
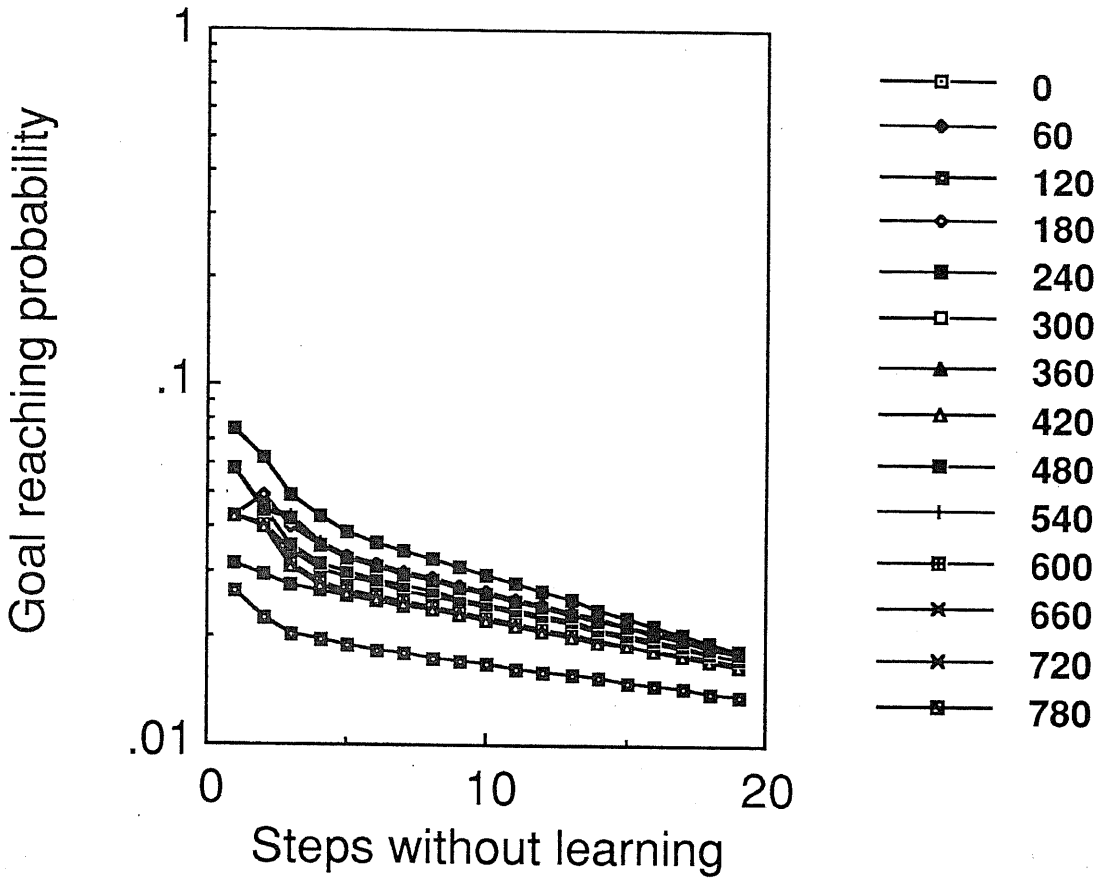


Fig. 4-32 確定的な環境における慎重な動作主体 (#2)
(報酬性ゴールのみを持つ環境)

上図：学習を止めた状態で報酬性ゴールに到達する確率。
下図：学習に伴う動作主体の状態変化など。



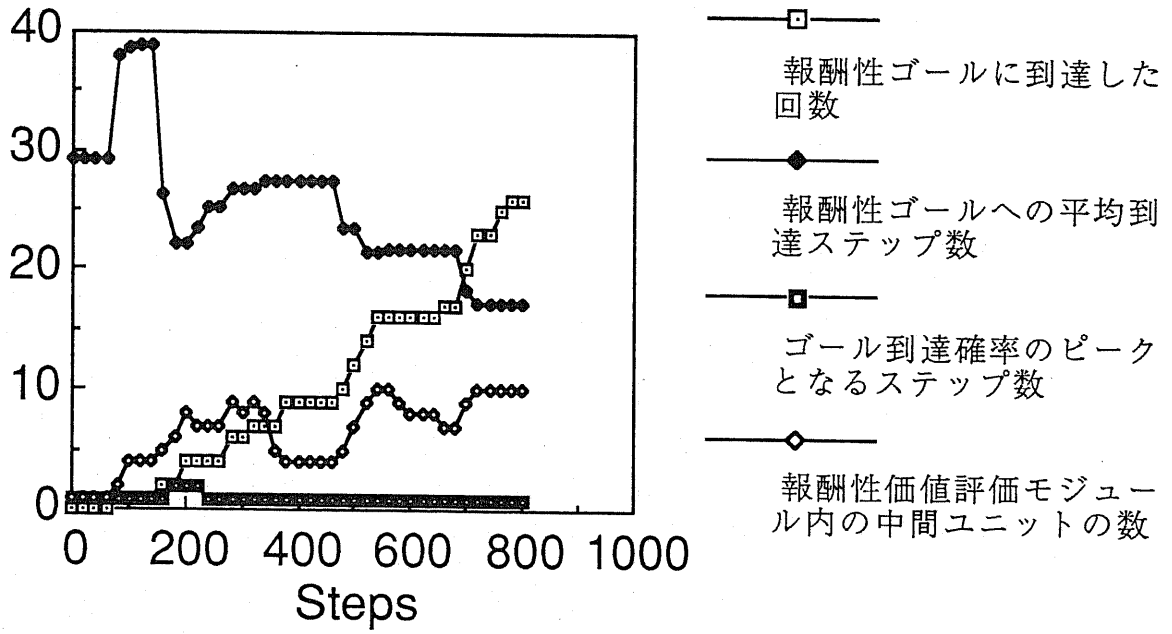
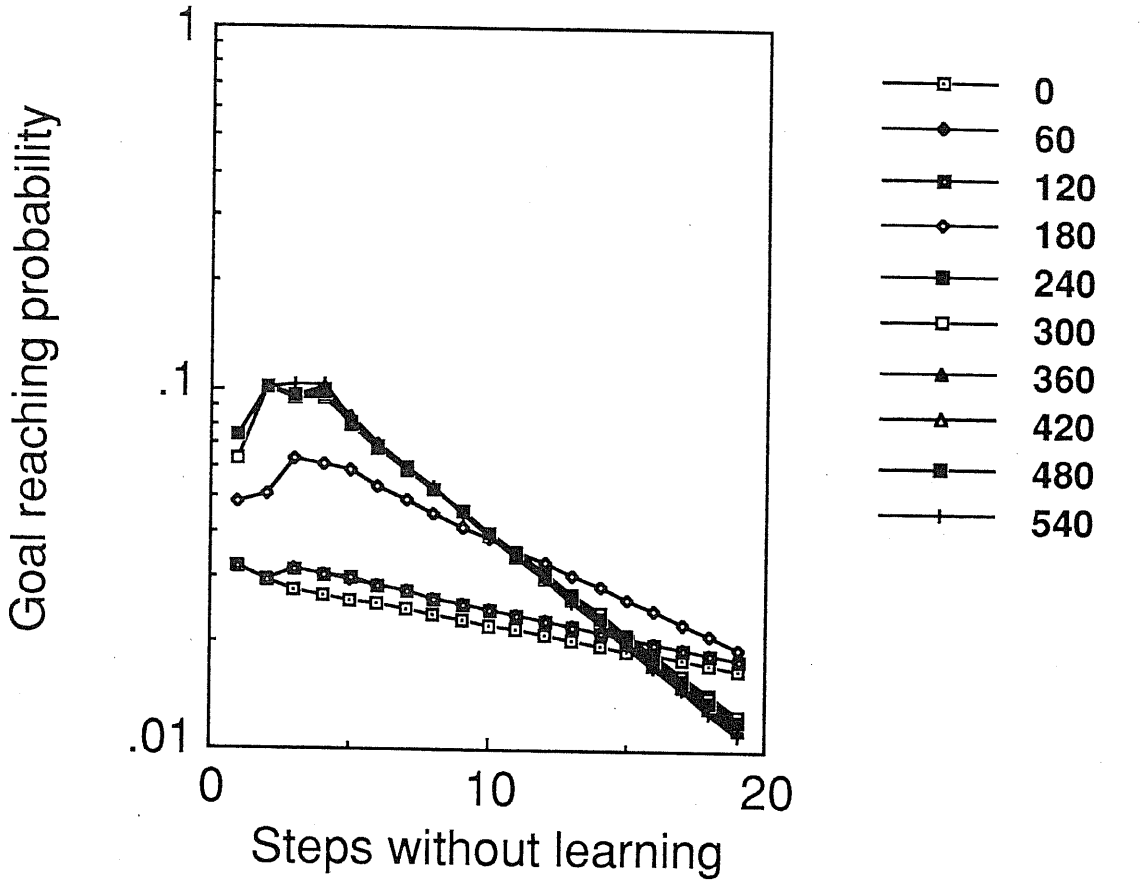


Fig. 4-33 確率的な環境における慎重でない動作主体 (#1)
(報酬性ゴールのみを持つ環境)

上図：学習を止めた状態で報酬性ゴールに到達する確率。
下図：学習に伴う動作主体の状態変化など。



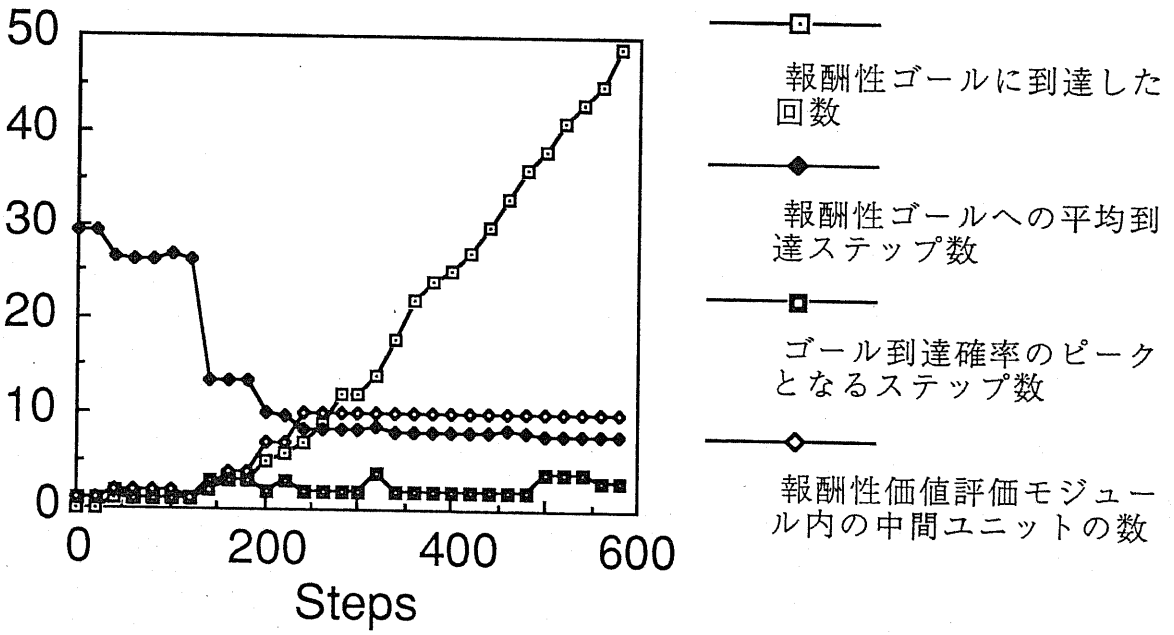
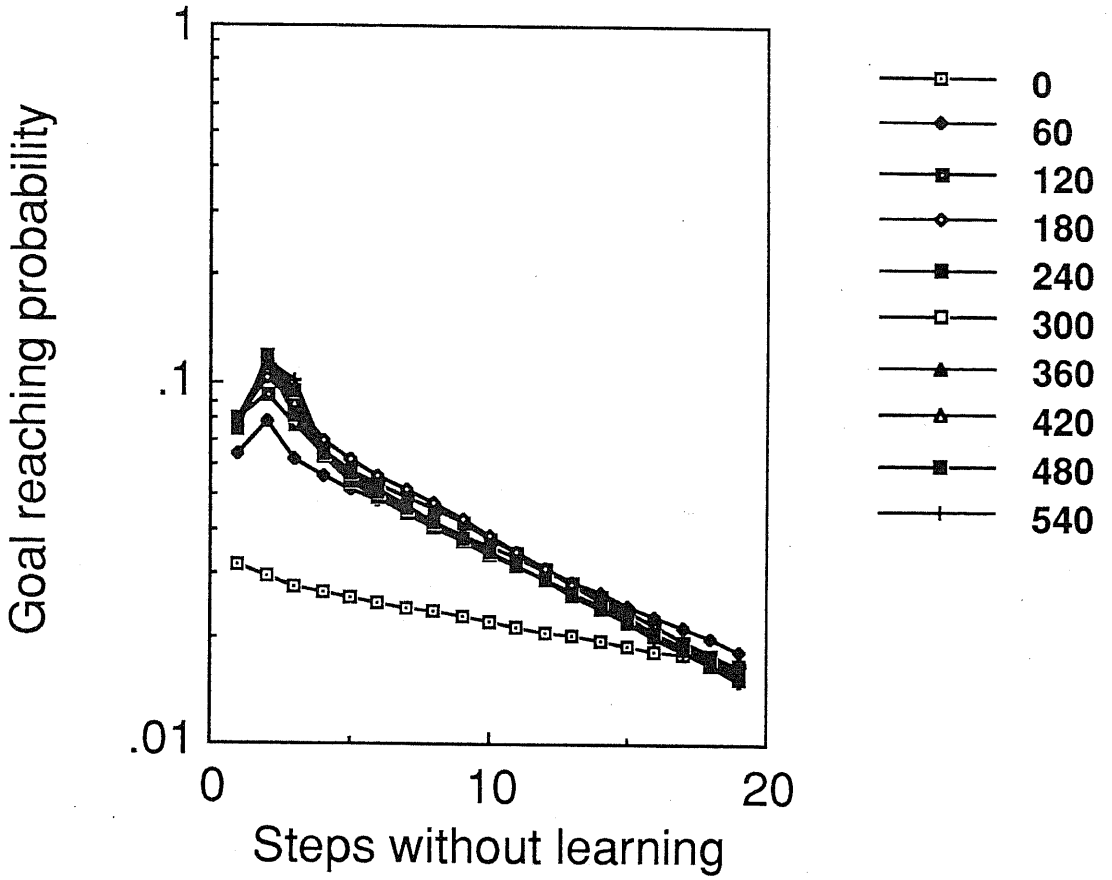


Fig. 4-34 確率的な環境における慎重でない動作主体 (#2)
(報酬性ゴールのみを持つ環境)

上図：学習を止めた状態で報酬性ゴールに到達する確率。
下図：学習に伴う動作主体の状態変化など。



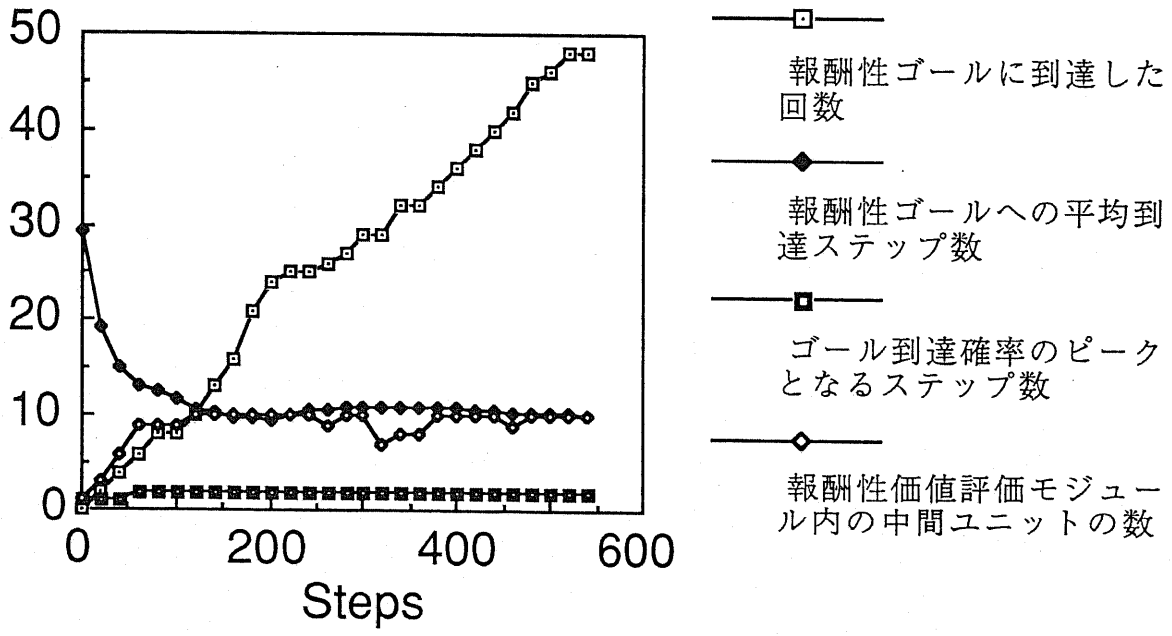
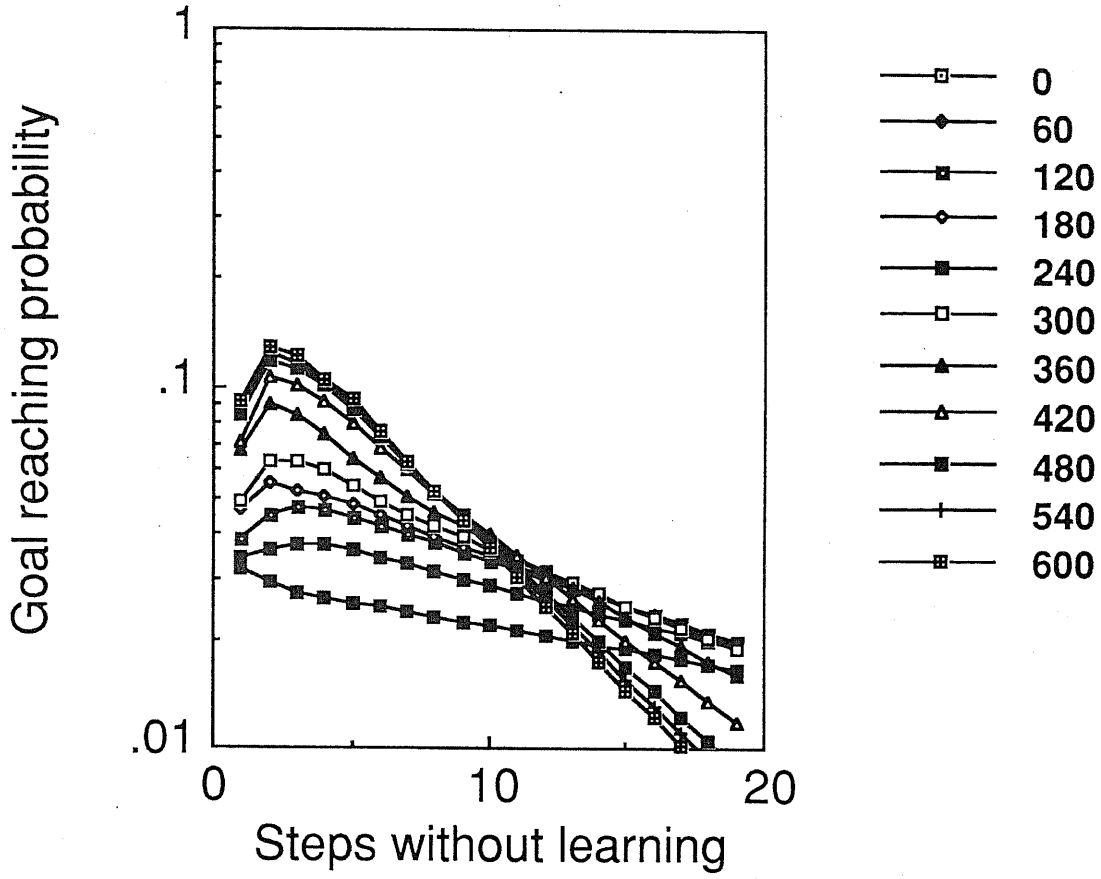


Fig. 4-35 確率的な環境における慎重でない動作主体 (#3)
(報酬性ゴールのみを持つ環境)

上図：学習を止めた状態で報酬性ゴールに到達する確率。
下図：学習に伴う動作主体の状態変化など。



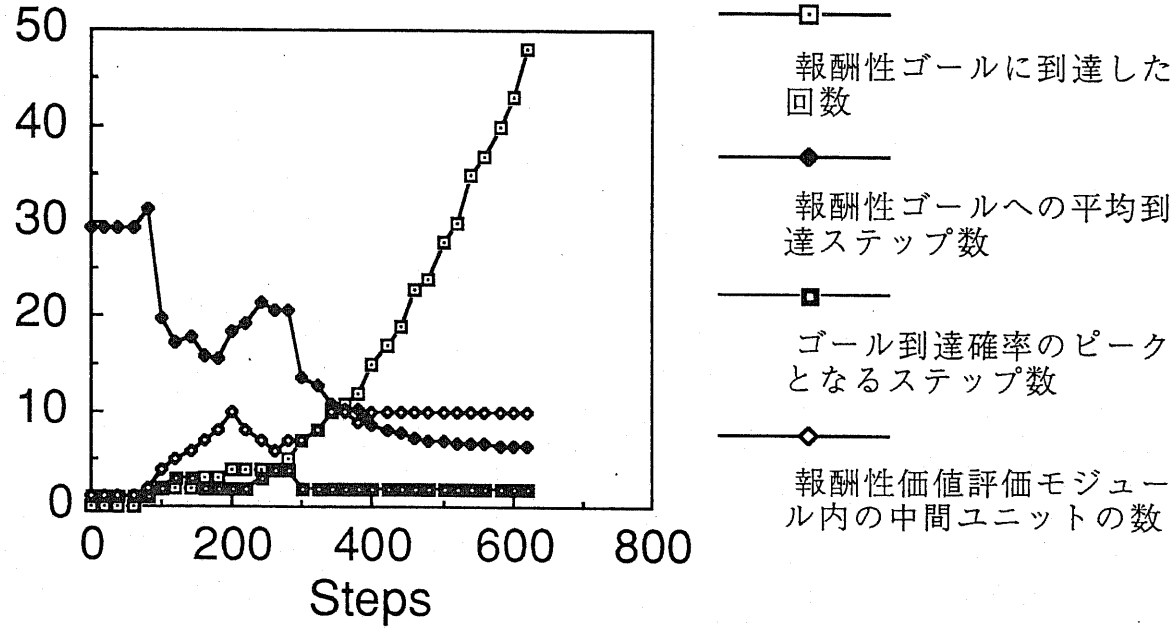
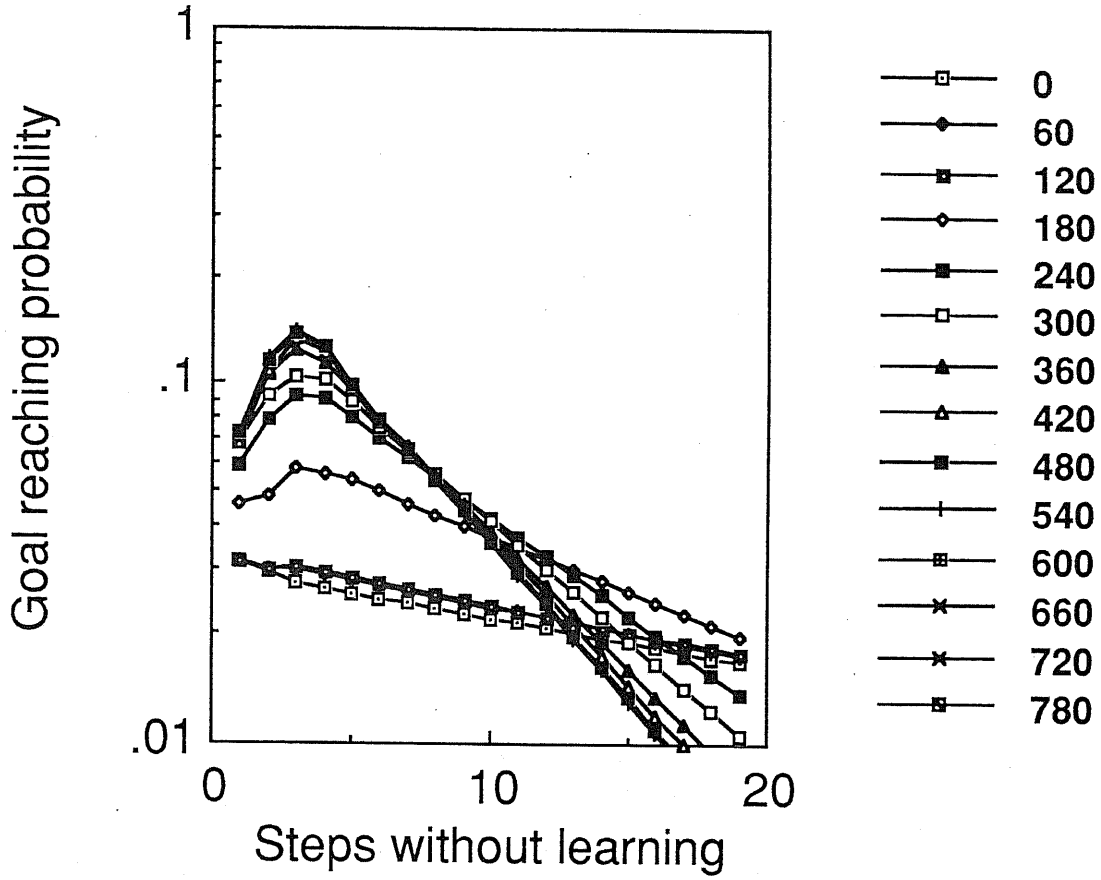


Fig. 4-36 確率的な環境における慎重な動作主体 (#1)
(報酬性ゴールのみを持つ環境)

上図：学習を止めた状態で報酬性ゴールに到達する確率。
下図：学習に伴う動作主体の状態変化など。



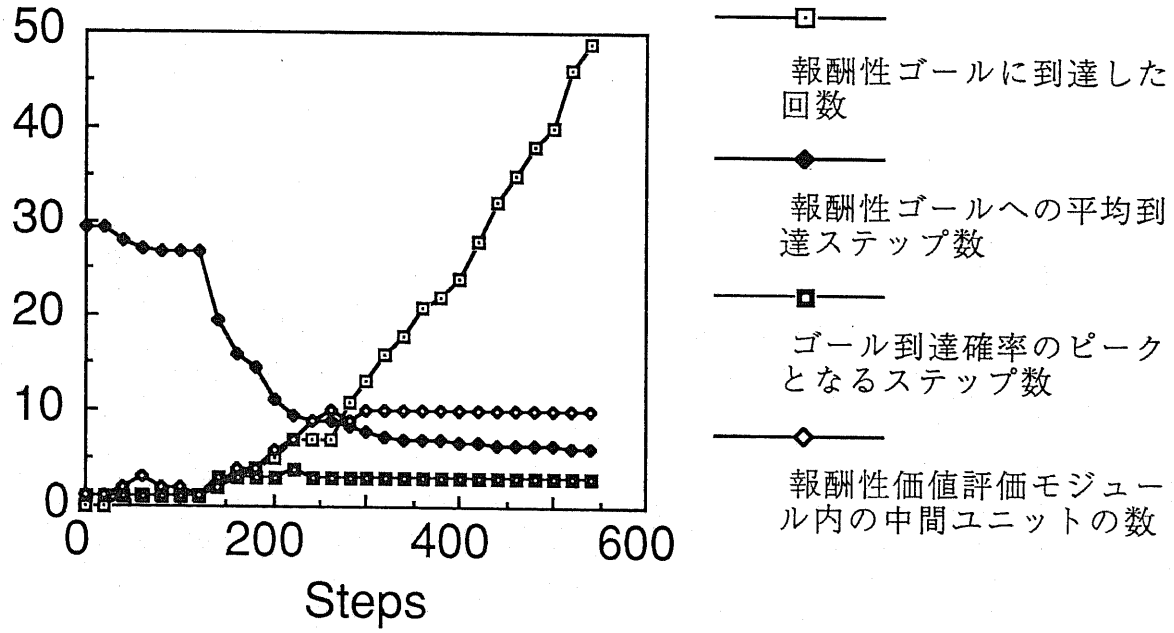
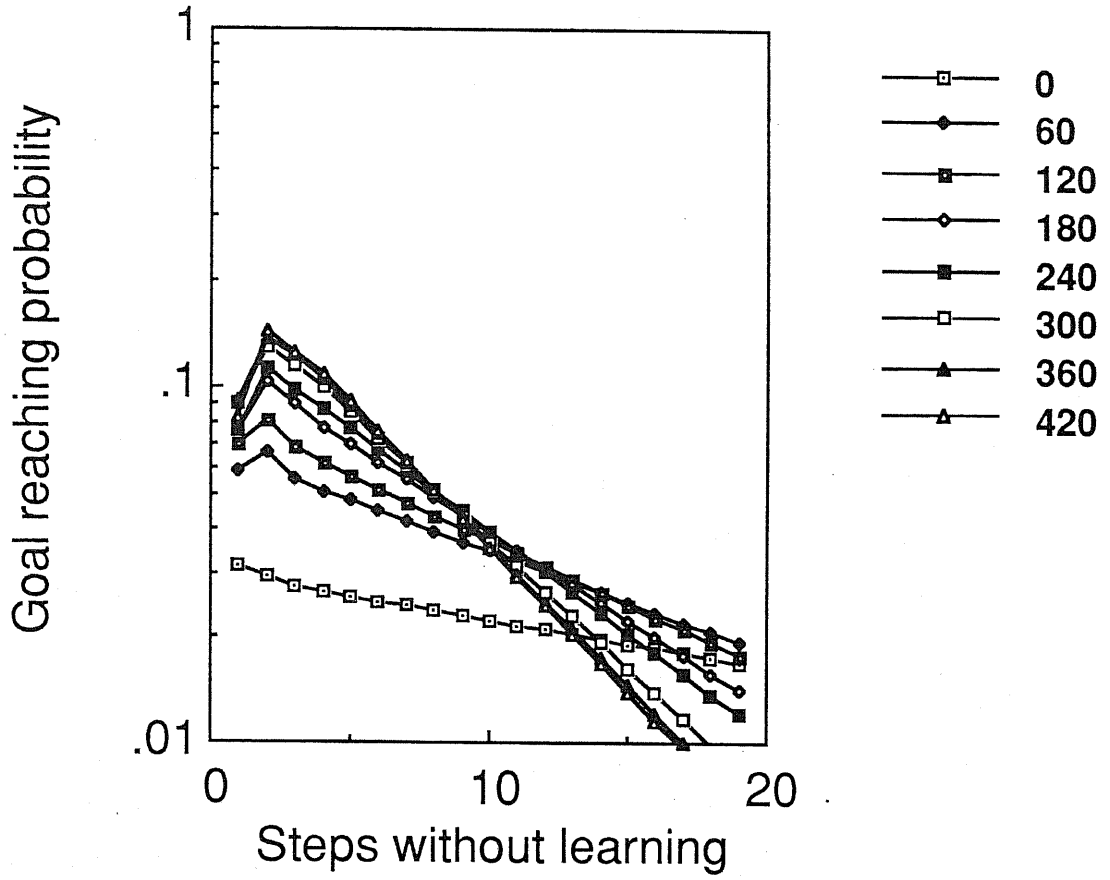


Fig. 4-37確率的な環境における慎重な動作主体 (#2)
(報酬性ゴールのみを持つ環境)

上図：学習を止めた状態で報酬性ゴールに到達する確率。
下図：学習に伴う動作主体の状態変化など。



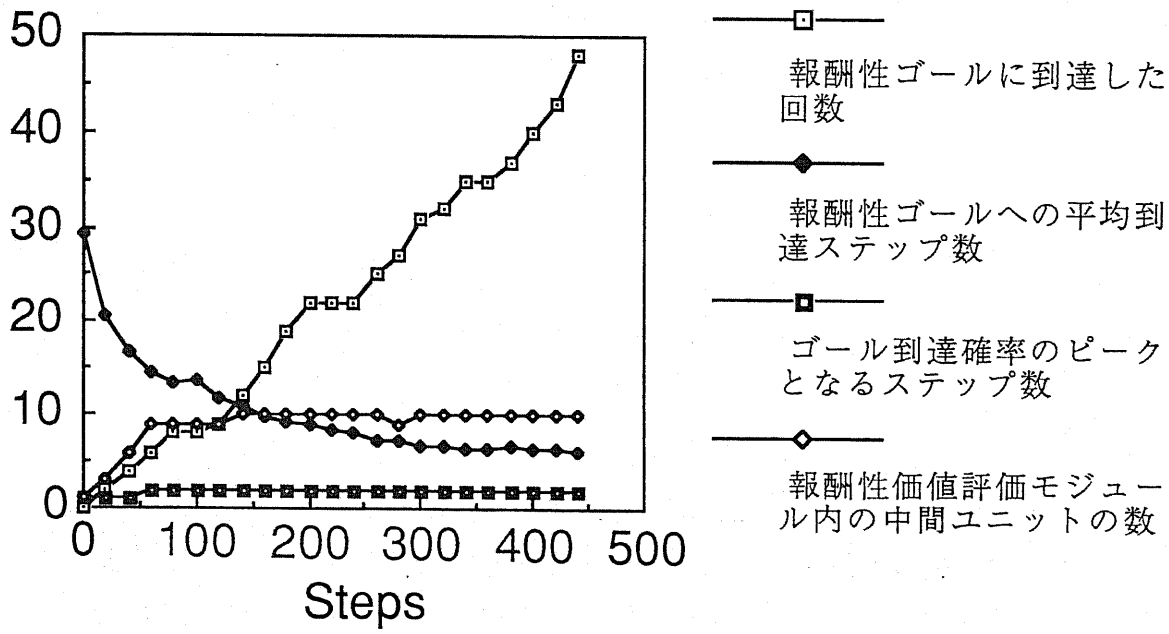


Fig. 4-38 確率的な環境における慎重な動作主体 (#3)
(報酬性ゴールのみを持つ環境)

上図：学習を止めた状態で報酬性ゴールに到達する確率。
下図：学習に伴う動作主体の状態変化など。

総　括

本研究では第１章で述べたようなより人間に近い知性を実現するために『自発的に思考するシステム』の開発をテーマとして研究を進めてきた。生物は環境と相互に作用を及ぼし合いながらその過程で適応を行ない更に知性を構築することから、本研究においても環境を含む全体のシステムを取り扱った。また現実の環境と同様に動作主体に対してあからさまに行動の修正を指示する教師信号を持たないこや、研究対象と成る動作主体が生体において原理的に実現可能であること等を要請した。そしてニューラルネットワークの動作原理をその基本的コンセプトに据えて研究を行なった。

第２章では思考をパターン情報に対する相関の強調であると仮定し、意識による作用の重要性を指摘した。次に、生物における環境に対する動作主体の適応を考慮し、３つのモジュールにより構成される動作主体のモデルを提案した。３つのモジュールはそれぞれ行動、世界観、価値観に相当する。価値観を構成するには動作主体が先見的に持っている基本的な価値を再帰的に拡張すれば良いことを示した。そして価値信号は通常の評価関数としての機能とあわせて、処理する価値があるかどうかの基準となる重要性を表現するので、大容量のパターン情報に対して注意点を与えるのに役立つ。このシステム自身が与えることができる注意点が人間における意識に対応するものだと考えている。つまり、世界観と価値観が意識を通して行なう相互作用が思考であると考えた。

第３章では上記システムを神経回路で実現するために必要な各モジュールにおける能力の検討から、認識連合モジュールにおいて一次記憶を実現することが重要な課題であると考えた。そのためには相互結合型ニューラルネットワークの活動状態を、パターン情報を表現している活動状態に制御すべきである。そこで情報処理を行なう興奮性ネットワークに対し抑制性ネットワークを付加した二層構造を用いて活動状態をダイナミックに制御する方法を提案した。さらに、興奮性ユニット毎に非線形な自己帰還入力を導入し、ネットワークの活動状態をより細かく制御することを可能とした。

第４章では、第２章で開発した環境を含むシステムを計算機シミュレーションにより実現することを目指した。ここでは価値観の形成に研究の主眼を置き、簡単化のために出力ニューラルネットワーク（行動）と価値評価モジュール（価値観）の二つのみを含む動作

主体が環境に適応する様子を調べた。第２章における基本モデルから再帰モデルに至る価値評価モジュールの改良に伴う能力の向上が確認され、同時に価値評価モジュール内では価値体系が形成されることが示された。さらに嫌悪性の価値に対する回避行動の学習や不確定性を含む確率的な環境における適応能力もシミュレーションにより確認した。

結局、本論文において為された主な成果は次のものである。１つは環境に適応しながら自発的に思考するシステムの全体像を与えたこと。２つ目に、その部分システムが既にある程度の適応能力を持つことを計算機シミュレーションで検証したこと。３つ目に、このシステムが神経回路によって実現できる可能性が強いことを示したことである。これらの成果は当初の目的に鑑みればまだまだ不十分であるが、自発的に思考するシステムの基礎は確立された。

パターン処理型の知能マシンに対する全体像が３つのモジュールによるシステムとして与えられたので、この枠組を用いれば認識能力を評価できる可能性がある。なぜなら行動を含めた研究では、行動能力により動作主体の評価ができるので、認識機構を変化させたいくつかの実験から認識機構の能力を評価することができるからである。さらに個々のモジュールの果たすべき機能がより明確になるために、各モジュール毎に段階的な研究を行なうことが可能になる。

また、本研究を通して次第に明らかになってきたことは、注意信号の存在が非常に重要だと言うことである。例えば、犬に投げたボールを拾ってくる課題を調教するときなどに報酬や罰を与えることは勿論であるが、同時に犬の注意を引くためにボールが目立つようにすることが重要である。また学生が学校で授業を受ける時にも、習う当人がその内容に適切な注意を向けていなければ効果は上がらない。この場合は前者と異なり注意は動作主体自身で向けなければならず、これが意識の作用に対応すると思われる。だから、集中力の続かない人というのは注意の固定が不得意な人だと思われる。そこで“批判信号と注意信号を用いた学習”が、現実世界に対応する適切な学習パラダイムではないかと提案する。このことから優れた学習援助システム（教師等）は各瞬間における注意のポイントを明確に指示する必要があるだろう。

さらに、第４章における価値観の構成に関する研究では一般化能力を無視するための方法として用いた一般化ができない環境はシンボルとパターンが区別できなくなる極限的状況であると考えられる。一般化能力が無視できれば、それ以外の知的システムに必要とされる能力を明確ににする研究を行ない易い。しかし逆に今後の発展的研究の過程では、我々の提案したモデルの各モジュールにおいても基本的には一般化能力等が必要である。けれ

ども価値評価モジュールにおいてはその前段にある認識連合モジュールがそれらの能力をを請け負うことができるので、必ずしも一般化能力などを必要とはしない。よって、本論文で主に研究を行なった価値観の形成に関しては、比較的問題の分離が容易であった。

将来的な発展としては、脳の構成と比較してさらなる知見を得る事や、心理学的な現象との比較も興味深い。特に神経言語プログラミングの研究(ﾊﾞﾝﾄﾞﾗｰ, 1986)などとは良い対応がとれそうである。さらにペットのように飼い慣らす事ができる機械を作りたいと考えている。人は身近なものほど細かく見えるので人間と他の動物の差がとても大きいと思い勝ちだが、生命の進化において多細胞化より後の進化がそれ以前の単細胞生物の完成にかかった時間に比べてはるかに短かったように(ﾏｰｸﾞﾘｽ, 1989)、犬や魚程度の知能でも実現できれば、その後のさらに人間に近付けるための改良は比較的容易だと思われる。よって、それら動物レベルの人工的な知能マシンの開発が、今後の一つの研究目標となり得ると思われる。

REFFERENCE

Barto A. G., Sutton R. S., & Brouwer P. S., "Associative search network: A reinforcemnet learning associative memory", Biol. Cybern., vol.40, pp.201-211, 1981.

Barto A. G., & Sutton R. S., "Simulation of anticipatory responce in classc diigy neuron-like elemint", Behev. Brain Res., vol.4, pp.221-235, 1982.

Barto A. G., Sutton R. S., & Anderson C. W., "Neuronlike adaptive elements that can solve difficult learning control problem", IEEE Trans. **S M C**-13, pp.835-846, 1983.

Barto A. G., "Learning by statistical cooperation of self-interested neuron-like computing elements", Human Neurobiol., vol.4, pp.229-256, 1985.

Blazis D. E. J., Desmond J. E., Moore J. W. & Berthier N. E., "Simulation of the classically conditioned nictitating membrane response by a neuron-like adaptive element: A real-time variant of the Sutton-Barto model", Proc. of 8th Annual Conf. of the Cog. Sci. Soc., pp176-186, .

Fukaya M., Kitagawa M., Okabe Y., "Two-level neural networks: Learning by interaction with environment", Proc. of IEEE 1st Annual Int. Conf. on Neural Networks, 87TH0197-7, pp.II531-II539, Jun. 1987.

Jameson J., "A neuroncontroller based on model feed back and the adaptive heuristic critic", Proc. of Int. Joint Conf. on Neural Networks, San Diego, vol.2, pp.37-44, 1990.

Klopf, A. H., "A drive-reinforcement model of single neuron function: An alternative to the hebbian neuronal model", AIP Conference Proceedings 151: Neural networks for computing, pp.265-270, 1986.

Klopf A. H., "Drive-reinforcement learning: A real-time learning mechanism for unsupervise

learning", Proceedings of the IEEE First Annual Int. Conf. on Neural Networks, San Diego, California, 21-24 June, vol.2, pp.441-445, 1987.

Klopf A. H., & Morgan J. S., "The role of time in natural intelligence: implications for neural network and artficial intelligence reseach", Proc. of Int. Joint Conf. on Neural Networks, Washington D. C., vol.2, pp.97-100, 1989.

Millan J. R. & Torras C., "Learning to avoid obstacles through reinforcement", Machine Learning, Proc. of the 8th Int. Workshop, pp.298-302, 1991.

Morgan J. S., Patterson E. C. & Klopf A. H., "A drive-reinforcement neural network model of simple instrumental conditioning", Proc. of Int. Joint Conf. on Neural Networks, San Diego, vol.2, pp.227-232, 1990.

Muro P., "A dual back-propagation scheme for scalar reward learning", Proc. of 9th Annual Conf. of Cognition, vol.I, pp.165-176, 1986.

Okabe Y., "Moderationism: Feedback learning of neural networks", Proc. of IECON, Oct pp.1028-1033, 1988.

Okabe Y., Fukaya M., Kitagawa M., "AND-OR logic analog of neuron networks", Intl. Conf. on Computer Simulation in Brain Science, Aug. 1986.

Pavlov I. P., Conditioning reflexes, Oxford Univ. Press, 1927.

Schmidhuber J., "Recurrent network adjusted by adaptive critics", Proc. of Int. Joint Conf. on Neural Networks, Washington, vol.1, pp.719-722, 1990

Schmidhuber J., "An on-line algorithm for dynamic reinforcement learning and planning in reactive environments", Proc. of Int. Joint Conf. on Neural Networks, San Diego, vol.2, pp.253-258, 1990

Sutton R. S., & Barto A. G., "Toward a modern theory of adaptive networks: Expectation and prediction", Psychological Rev., vol.88, No.2, pp.135-170, 1981.

Sutton R. S., "Learning to predict by method of temporal differences", Machine Learning, no.3, pp.9-44, 1988.

Tesauro G., "Simple neural models of classical conditioning", Biol. Cybern., vol.55, pp.187-200, 1986.

Werbos P., "Building and understanging adaptive systems: A statistical/numerical approach to factory automation and brain research", IEEE Trans. SMC-17, no.1, pp.7-20, 1987.

Werbos P., "Backpropagation and neurocontrol: A review and prospectus", Proc. of Int. Joint Conf. on Neural Networks, Washington D.C., vol.1, pp.209-216, 1989.

Widrow B., Gupta N. K., & Maitra S., "Punish/Reward: Learning with a critic in adaptive threshold systems", IEEE Trans. SMC-3, pp.455-465, 1973.

Williams R. J., & Peng J., "Reinforcement learning algorithms as function optimi zers", Proc. of Int. Joint Conf. on Neural Networks, Washington D.C., vol.2, pp.89-95, 1989.

合原一幸, "ニューロコンピュータの将来像", コンピュートロール no.24, pp.126-131, 1988.

甘利俊一, 神経回路網の数理, 産業図書, 1978.

伊藤正男, ニューロンの生理学, 現代科学選書, 1972.

伊藤正男, "大脳と小脳", 生体の科学, vol.40 no.2, Mar.-Apr., pp.82-89, 1989.

岩井栄一, 渡辺譲二, 靱負正雄, "側頭葉と視覚性学習・記憶", 生物の科学　遺伝　別冊 no.2 1989-11, pp.56-67.

大輪勤, "モデレーショニズムに基づくフィードバック学習", 修士論文, 1989.

小野武年, 福田正治, "海馬と学習・記憶", 生物の科学　遺伝　別冊 no.2 1989-11, pp.68-77.

小野武年, "扁桃腺体における生物学的価値判断の機構", 国際フロンティアシンポジウム フロンティア研究 脳の思考機能へのアプローチ 講演要旨, pp.14, 1990.

小林繁, 熊倉鴻之助,黒田洋一朗 ,畠中寛, 絵ときブレインサイエンス入門, オーム社

坂本邦博, "外界と相互作用を行ないながら自己組織化する神経回路の研究", 修士論文, 1983.

佐賀一繁, 菅坂玉美, 関口実, 長田茂美, "自己学習方式の一検討", 1989年電子情報通信学会秋期全国大会, 1989.

ｽｸｧｲｱｰ L. R., 記憶と脳 -心理学と神経科学の統合-, 医学書院,1989. (Squire L. R., Memory and brain, Oxford univ. press, 1987.)

ﾄﾞｰｷﾝｽ R., 利己的な遺伝子, 紀ノ国屋書店, 1991. (Dawkins R., The selfish gene, Oxford univ. press, 1989.)

銅谷賢治, "ニューラルネットワークの振動パターンの記憶", コンピュートロール no.29, pp.52-62, 1990.

中野馨他, 入門と実習 ニューロコンピュータ, 技術評論社, 1989.

中野馨, 合目的行動を自己形成するロボット, 東京大学工学部計数工学科 授業　自己組織システム論 11, 1987.

ﾊﾞﾝﾄﾞﾗｰ R., 神経言語ﾌﾟﾛｸﾞﾗﾐﾝｸﾞ : 頭脳をつかえば自分も変わる, 東京図書, 1986. (Bandler, R., Using your brain : for a change,,.)

深谷正道, "モデレーショニズム -外界との相互作用により学習する神経回路-", 修士論文, 1988.

福島邦彦, "視覚パターン認識のモデル", 生物の科学　遺伝　別冊 no.2 1989-11, pp.102-110.

ﾏｰｸﾞﾘｽ L., & ｻｶﾞﾝ D., ミクロコスモス -生命と進化-, 東京化学同人, 1989. (Margulis L., &

Sagan D., Microcosmos,,.)

宮下保司, "大脳のメモリーニューロン", 生体の科学 vol.40 no.2, Mar.-Apr., pp.121-126, 1989.

ﾘﾝｽｷｰ R., "知覚神経回路における自己組織化", 人工ニューラルシステム, vol.21, no.11 ,pp.71-86, 1989. (Linsker R., "Self-organization in a perceptual network", IEEE Computer, vol.21, no.3, pp.105-117, 1988.

発表文献等

山川宏, 岡部洋一, "相互結合を持つネットワークのスパース安定化", 神経回路学会平成2年 全国大会講演論文集, P2-21, pp.84 Sep 1990.

山川宏, 岡部洋一, "相互結合を持つネットワークの情報処理可能性", (株)富士通 KSAﾌｫｰﾗﾑ・ﾆｭｰﾛｺﾝﾋﾟｭｰﾀ分科会講演資料, Sep 1990.

山川宏, 岡部洋一, "逐次的に現れるパターンを効率的にニューラルネットワークに学習させる方法", 1991年電子情報通信学会春期全国大会, D-23, 1991.

Yamakawa H., & Okabe Y., "Self-organization of secondary values on neural automaton", Int. Joint Conf. on Neural Networks, Seattle, Supplementary poster program SW63, pp.II A-988, 1991.

山川宏, 岡部洋一, "強化学習に基づく知能システム", (株)富士通 KSAﾌｫｰﾗﾑ・ﾆｭｰﾛｺﾝﾋﾟｭｰﾀ分科会講演資料, Sep 1991.

Yamakawa H., & Okabe Y., "A recursive neural system for memorizing systems of values arranged in a tree like structure", Int. Joint Conf. on Neural Networks, Singapore, vol. 2, pp. 1776-1781, 1991.

山川宏, 岡部洋一, "強化学習に基づく知能システム", , 神経回路学会平成3年 全国大会講演論文集, P3-4, pp.125-126, Sep 1991.

Y amakawa H., & Okabe Y., "Intelligent system based on reinforcement learning", Neural Networks. (投稿予定)

謝　　辞

今年もまたスキーのシーズンがやってきて長かった学生生活もおわりに近づいてきたようです。振り返れば3年前の今ごろ修士論文を書くために実験に追われていた事を思い出します。修士まで物理をやっていた私が博士課程から突然ニューラルネットワークの研究に転向したにもかかわらず、ここまで何とかやってこられたのは、周りの皆様の暖かいご援助のたまものであると考えています。

まず担当指導教官である岡部洋一教授には大変自由な環境の上に度々適切な御指導を頂き感謝しております。また蔵王のワークショップでは東京大学の甘利俊一教授、吉沢教授、理化学研究所の高橋哲也氏に研究にとって重要な方向性を示して頂きました、さらにATRの川人光男氏には参考文献等を教えていただき大変役立ちました。銅谷賢治氏には参考文献の収集等で力添えを頂きました。塚田稔教授には生物的な知識を丁寧に教えていただきました。大森隆司氏にはネットワーク上でのお話などを含め幾度かにわたりご批評を頂きました。

2度にわたるKSAフォーラムでは浅川和男氏、長田茂美氏、松尾和洋氏をはじめとする富士通の方々にたいへん貴重なご意見を頂き参考になりました。また益岡竜介氏には岡部研究室のSUNワークステーション及びユニックスの取扱いについて細かくご指導頂きました。木本氏には参考文献等を紹介して頂いたほかにシンガポールでは楽ませて頂きました。

研究室においても、広瀬明氏には打ち合わせ等で厳しいご批判を頂きました。北川学氏及び宮尾光生氏には日々の物品の購入に辺り親切なご指導を頂きました。中山秋芳氏にはデバイス研究の視点から色々とご意見を伺いました。田宮寿美子氏にはいつも暖かく声を頂きさらにしばしば夜食を持ってきて頂き、また小島潤子氏や二年間お世話になった鈴木明子氏には旅行手続きでたいへんお世話になった上にいつもお茶をいれて頂きました。

松原玄宗氏には研究室や本郷のある1室で度々鋭いご指摘を頂き、金氏からは日々の生活に潤いをあたえる楽しいお話を何度となく伺いました。大豆生田利章氏にはアイドルにまつわる楽しい話をたくさん伺いました。甲原隆矢氏とは研究内容に関して議論をする機会を持てた事を感謝します。またパブロ氏とはシンガポールでの口頭発表の準備でお世話になり、タイでの観光では楽しい一時を過ごさせて頂きました。鈴木晃治郎氏にはネットワークシステムの構築にご助力頂いた上に、テニスコートでも厳しいご指導を頂きました。宮崎祐行氏には研究室旅行の準備や力仕事を通して大変厄介になり、コンピュータのネーミングも手伝って頂きました。。フランク氏には英語論文の手直し等を通してお世話になりました。おそらく社会人に復帰した松井俊之氏には酒の席で楽しい話を度々聞かせて頂きました。木村浩氏には研究と現在の理科大の状況を合わせた大変楽しいお話を伺う事ができました。さらに、木村勝氏にはツモに酒にと楽しませて頂きました。短い間ではありましたが住友電工の徳田人基氏にも打ち合わせにおいて意見を伺う事ができました。

なお中野馨氏には修士課程時の授業でニューラルネットワークに興味を持つ契機を与えて頂いた他にKSAフォーラムでも考えをお聞きできました。

来年度からは富士通の川崎研究所で引き続きニューラルネットワークの研究を続けて行くつもりですので、これまでの皆様のご援助に答えられるよう、より一層の努力をして行きたいと考えています。

ここで以上の皆様に感謝の意を表し、謝辞とさせて頂きます。