強化学習による探索行動の学習
Learning of Exploration Behavior by Reinforcement Learning

柴田克成
大分大学工学部 〒870-1192 大分市大字旦野原 700 番地
Katsunari Shibata
Oita University, 700 Dannoharu, Oita 870-1192
shibata@cc.oita-u.ac.jp

Abstract

Exploration is an important factor that influences the performance in reinforcement learning, and random actions are usually employed as the exploration. However, the exploration that real lives are doing does not seem only a sequence of random actions, but seems a kind of deterministic and intelligent actions in which the context is considered. In this paper, the author tries to explain such explorations as a deterministic behavior and propounds a novel approach that exploration is acquired by reinforcement learning. Then, it is shown that an agent with a recurrent neural network that is trained based on reinforcement learning becomes to explore in some degree through learning in two simple problems.

Keywords: exploration, reinforcement learning, recurrent neural network, context

１. はじめに

強化学習は、試行錯誤をすることにより、他の助けを必要とせず、報酬や罰から自ら学習することができる強力な合目的的かつ自律的な学習則である[1]。この試行錯誤は探索(exploration)と呼ばれ、強化学習のパフォーマンスを左右する一つの重要なファクターとなっている。通常は乱数を用いた確率的な行動選択を行うことでそれを実現するが、代表的なものに、ε-greedy やボルツマン選択などが挙げられる[1]。このとき、行動選択の確率的要素が大き過ぎると、その時点で最適でない行動をとる機会が増え、パフォーマンスの低下につながる。一方、確率的要素が小さ過ぎると、その時点でのパフォーマンスは良くなるが、それよりも良い行動があっても、なかなかそれを発見できないため、将来的なパフォーマンスを考えると得策ではないという「exploration と exploitation のジレンマ」の問題が指摘されてきた。そこで、確率的要素を、最初は大きくして、その後、徐々に小さくしたり、各状態の行動価値関数の行動に関する分散に応じて変化させたり[2]といった手法が提案されるなど、確率的要素をいかに制御するかも大きな問題となってきた。

筆者は、この強化学習が、行動の学習だけでなく、認識や記憶なども含めたあらゆる機能の学習に有効であり、われわれ生物の知能を説明する上で非常に重要な役割を果たしていると考えてきた[3]。本論文の内容は、われわれ生物の脳の中で強化学習が行われていると考えたときに、そもそも「われわれ生物は確率的な行動をとっているのであろうか？」「われわれ生物の脳の中には確率的な要素を生み出すための乱数発生器があるのだろうか？」という素朴な疑問を出発点とする。たとえば、われわれが初めて経験する迷路の中に入った場合でも、過去に通ったところは通らないとか、さらには、２次元のマップを頭に思い描くなどして、知識を動員し、文脈を考慮して探索を行う。また、分かれ道に差し掛かったときも、その前で指を動かしてみたり、ジャンプしてみたり、分かれ道の真ん中を行ってみるという探索は通常はしないということを考えると、「探索」は単なる確率的行動というよりは、逆に非常に知的な行動に見える。また、どちらに行ったら良いか全く分からない場合、サイコロを振るという決断を下すことはあるが、決断自身は確率的ではなく、決定論的に行っていると考えることもできる。

これに対し、「知識や文脈を用いてより適切な行動価値を表現し、その中でより高い行動価値が得られる行動を高い確率で選択している結果である」という説明や、「各アクチュエータレベルでの探索は行っていないが、もっとより抽象化された行動空間において確率的な行動選択を行っている結果である」といった説明は可能であろう。そして実際に、行動空間の抽象化に関する議論は、時間軸方向の抽象化(temporal abstraction)として議論されているところである[4]。

確かに、確率的行動を行っている可能性を否定することは難しいが、逆に、「確率的要素なしで未知の領域をくまなく探索できるだろうかという心配」と、「統計

的な議論が容易になる」ということ以外には、確率的要素を用いる必然性も見当たらない。そして前述のわれわれ自身の「探索」を振り返ってみると、確率的要素を用いなくても探索はできるのではないかと考えられる。探索が何らかの決定論的な意志決定の結果であるとすれば、これは広い意味で行動の一種と考えることができる。したがって、強化学習によって、他の機能と同様に、より良い探索も実現できるのではないかと期待される。

以上より、本論文では、われわれ生物が行っている「探索」は、確率的な行動選択ではなく、決定論的行動選択により実現されているのではないかとの考えに基づき、強化学習を適用することにより、学習によって「探索」と解釈できるような行動を獲得することを提唱する。ただし、学習を行うためにも「探索」が必要となるが、「学習のための乱数発生器を用いない探索」については、別途研究を進めているところである。したがって、本論文では、学習自体は「乱数による確率的行動決定」を用いて行い、学習した結果の行動として「乱数によらない探索行動」の獲得を目指す。また、前述のように、効率的な「探索」を行うためには、過去の文脈を有効に利用する必要があるため、リカレントニューラルネットを用いる[5]。そして、非常に簡単なタスクではあるが、シミュレーションの結果、ある程度効率的な探索行動を獲得できることを確認したので報告する。

２．探索とは何か？

前章では、探索を、乱数を用いた確率的行動決定として捉えるのではなく、通常の行動決定の枠組みで捉えることを述べた。では、通常の行動と探索は何が違うのであろうか。「探索」という言葉からは、通常は、明確なゴールが分からない状態での行動決定であるか、または、その行動によって直接ゴールに到達することよりも、その行動によって何らかの知識、例えば，迷路の構造を知るなどの知識を得ることによって、後の行動決定、そしてゴール到達に役立つことを意味すると考えることができる。そして、ここでは、単に乱数を用いて探索するよりも、より効率的な探索行動を学習によって獲得することを目指す。

そこで、本論文では、明確なゴールは分からないが、知識や文脈を用いた効率的な探索を行うという点に主眼を置いたタスクと、探索により知識を貯えるということに主眼を置いたタスクの二つを取り扱う。

３．学習方法

ここでは、最も一般的なリカレントニューラルネットとして、中間層ニューロンの出力を次の時刻の入力として扱う Elman 型のリカレントネットを用いた。そして、現在の観測値 s_t をニューラルネットへの入力とし、行動数と同じだけの出力ニューロンを用意し、その出力を各行動に対応するQ値とした。そして、Sarsa のアルゴリズム[1]に基づいて、１単位時間前の観測値を再度入力し、１単位時間前にとった行動に対するQ値の出力 $Q_{a_{t-1}}(s_{t-1})$ に対する教師信号 $Q_{s,a_{t-1}}$ を

$$Q_{s,a_{t-1}} = r_t + \gamma Q_{a_t}(s_t) \qquad (1)$$

r: 報酬、 γ: 割引率

と自動生成し、ニューラルネットの当該出力の部分のみ教師信号を与え、BPTT(Back Propagation Through Time)[6]に基づいて時間をさかのぼって教師あり学習させた。中間層、出力層の各ニューロンの出力関数は -0.5 から 0.5 の値域のシグモイド関数とした。

また、ここでは、学習はエピソード(試行)ごとに区切って行い、各試行の開始時には中間層からのフィードバック入力の値はすべて 0.0 とし、ゴール到達時以外のときは報酬 r を 0.0、ゴール到達時には報酬 r を 0.8 とした。また、ゴール到達時の状態におけるQ値を 0.0 とした。ニューラルネットの出力関数を 0.5 から-0.5 のシグモイド関数としているため、実際には、ニューラルネットの出力値に 0.4 を足した値をQ値とし、(1)式から求めた教師信号から 0.4 を引いた値を実際の教師信号としてニューラルネットに与えた。

４．シミュレーション

4.1 二者択一の探索行動

まず、はじめにで述べた分かれ道での探索において、どちらを選んだら良いか全く分からない状態で、まず、最初に片方を選択し、もしだめならもう一つを試してみるという形の探索が実現できるかどうかを検証した。

図１のように、5x5 のマス目の中央にエージェントを置き、４辺のどこかにランダムに２つのゴールの目印を置く。しかし、実際のゴールは２つの目印がある場所のうちのランダムに決めたどちらか片方だけとし、

エージェントはそれを事前に知ることができないとする。したがって、エージェントはまずどちらか片方のゴールの目印があるところに行き、それが本当のゴールでない場合は、もうひとつのゴールの目印の方へ行くことが要求される。そして、最初に２つのゴールの目印のうちのどちらかに向かうこと、そして、向かった先が本当のゴールではなかった場合に、もう一つの方のゴールに向かうことができるかどうかを確認する。

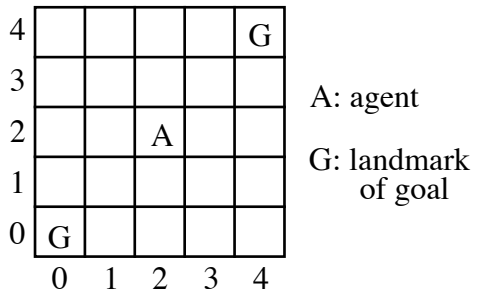

図１　二つのゴールの目印のうち、どちらが真のゴールであるかわからないタスク

ここでは、動作は上下左右への移動の 4 通りから選択でき、状態遷移は決定論的であるとした。壁にぶつかった場合は、その場にとどまることとした。また、観測値は、エージェントがどこにいてもゴールの目印が見えるように、自分を中心とした 9x9 の近傍の領域のマスのそれぞれに、ゴールの目印があるかどうかを示す０か１の信号 81 個で、それをニューラルネットにそのまま入力した。ニューラルネットは 3 層の Elman ネット、中間層には 20 個のニューロンを配し、BPTT でその試行の最初までまたは 30 ステップまでの短い方までさかのぼって学習した。確率的行動による探索は ε−greedy とし、ε は 0.1 で固定とし、割引率 γ は 0.92 とした。ニューラルネットの初期重み値は、中間層−出力層間はすべて 0.0 とし、入力層−中間層間は、セルフフィードバック結合を 4.0 それ以外のフィードバック結合を 0.0、外部入力に対する重み値を-0.5 から 0.5 の乱数とした。BPTT の学習係数は 0.2 とした。

学習開始当初は、２つのゴールの目印のうちの片方が先に学習されるため、行った先が真のゴールでなくても、なかなかそこを離れられず、１試行にかかるステップ数が増大し、真のゴールへ到達するまで数千ステップかかることもあったが、その後減少していった。

100000 回学習した後、4 辺のいずれかにゴールの目印を置いた場合の組み合わせ 256 通り（重なりを許す）について、エージェントに greedy に行動させた場合、５通りで失敗したものの、残りの 251 通りは、すべて真のゴールにたどり着くことができた。ゴールの目印を左下と右上の角に置き、本当のゴールを左下とした場合、ゴールの目印を左上と右下の角に置いて実際のゴールを右下とした場合、および、それらのゴールの位置を少しずらして、真のゴールを入れ替えた場合のエージェントの行動を図２に、また、最初の場合のゴール到達までの各Ｑ値の変化の様子を表１に示す。

図２を見ると、まず、エージェントは始めにどちらか片方のゴールの目印まで進み、それが真のゴールでないことを確認すると、もう一方のゴールに向かっていることがわかる。また、多少、行ったり来たりといった冗長な行動をとっている部分があるものの、全体として最適に近い行動を取っていること、さらに、図２の(c)(d)のように、(a)(b)の際に、後で行く方のゴールの目印をエージェントの方に一つ近づけると、エージェントは、近い方を先に行くようになっていることもわかる。ただし、遠い方のゴールの目印に先に行ってしまう場合も何回かはあった。

また、表１の１試行中のＱ値の変化を見ると、各状態での最大のＱ値は最初のゴールの目印に向かうにしたがって徐々に増大しているが、ゴールの目印に到着し、そこがゴールでないとわかるといったん減少する。しかし、次のゴールに近づくにしたがってＱ値は再び上昇している。また、この場合、右上(4,4)のゴールの目印に到達した後、到達前に通ったマス (3,4)(2,4)を再度通っている。ここでは、目標物の位置が変化しない限り、同じマスでは同じ観測値が得られる。にもかかわらず、右上(4,4)まで行く前と後では、各行動に対応するＱ値が大きく異なり、とった行動も右と左と逆向きになっていることがわかる。これは、最初のゴールの目印に到達したけれども報酬を得られなかったという情報をリカレントネットが何らかの形で保持し、それを元にＱ値を変化させているためと考えられる。

そこで、図３に、ある一つの中間層ニューロンの値が、図２の(a)の場合と(b)の場合で、時間とともにどう変化するかを示した。このニューロンは、最初のゴールの目印に到達するまでは 0.0 近傍の値をとっているが、いったんゴールの目印に到達して、そこが真のゴールでないとわかると、-0.4 程度の値を取るようになっている。これ以外のゴールの配置パターンの場合に

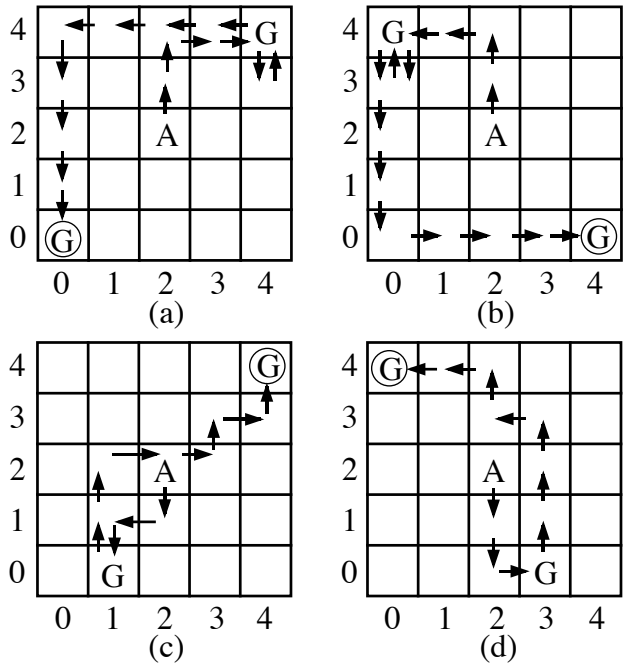

図２　学習後のエージェントの行動。G はゴールの目印を表し、○が付いた方が本当のゴールを表す。

表１　図２の(a)の場合の I 試行中の Q 値の変化

	上	右	下	左
(2, 2)−上−>(2, 3)	0. 49	0. 43	0. 48	0. 45
(2, 3)−上−>(2, 4)	0. 52	0. 49	0. 47	0. 45
(2, 4)−右−>(3, 4)	0. 51	0. 54	0. 45	0. 45
(3, 4)−右−>(4, 4)	0. 44	0. 56	0. 49	0. 49
(4, 4)−下−>(4, 3)	0. 32	0. 36	0. 39	0. 34
(4, 3)−上−>(4, 4)	0. 43	0. 37	0. 41	0. 39
(4, 4)−左−>(3, 4)	0. 34	0. 30	0. 34	0. 40
(3, 4)−左−>(2, 4)	0. 34	0. 33	0. 39	0. 46
(2, 4)−左−>(1, 4)	0. 40	0. 42	0. 47	0. 50
(1, 4)−左−>(0, 4)	0. 46	0. 51	0. 53	0. 54
(0, 4)−下−>(0, 3)	0. 44	0. 51	0. 55	0. 52
(0, 3)−下−>(0, 2)	0. 46	0. 47	0. 56	0. 52
(0, 2)−下−>(0, 1)	0. 49	0. 45	0. 57	0. 54
(0, 1)−下−>(0, 0)	0. 56	0. 51	0. 63	0. 60

ついても同様な傾向が見られることから、このニューロンは、いったんゴールの目印に到着したが真のゴールでなかったという情報を表現していると考えられる。

4.2 目標物が見えないときの探索

迷路に入れられたネズミは、エサがなくても探索するが、強化学習によって、この探索の行動を学習して

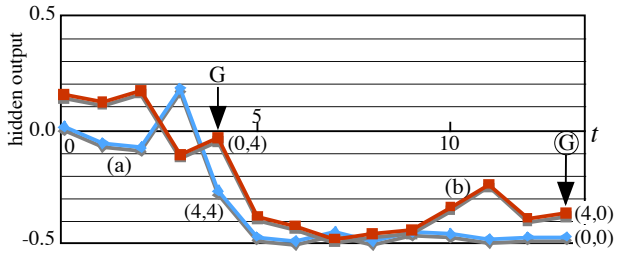

図３　図２の(a)(b)の際の、ニューラルネット中間層のある一つのニューロンの出力値の変化。G はゴールの目印、○が付いた方が本当のゴールを表す。

いると考えた場合、このような探索行動が学習によって獲得できるのかという疑問が湧く。そこで、次に、簡単な迷路の探索問題において、目標物が置かれる前にエージェントが迷路内に置かれ、事前に探索をして迷路の構造を把握しておくと、目標物が置かれた際により早く目標物に到達できる可能性が高くなるという設定で学習を行わせた。そして、目標物が置かれる前に、探索と呼べるような行動を行うことができるようになるかどうかを検証した。

まず、図４のように、2x2 のスペースに、一つの壁があるもの４つと壁のないもの１つの計５つの迷路を用意し、毎回ランダムに迷路を選ぶ。そして、エージェントを４個のマスの中の一つにランダムに配置する。

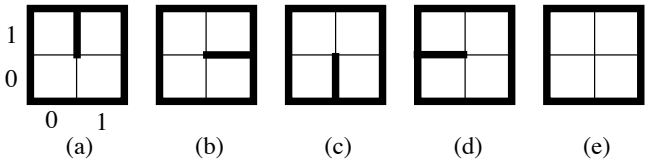

図４　シミュレーションで用いた５つの迷路

エージェントは、各ステップにおいて、上下左右の隣接するマスへの移動およびその場にとどまるという５つのうちの一つの行動をとり、状態遷移は決定論的とした。壁にぶつかる行動を選択した場合は、その場にとどまることとした。観測信号は、四方に壁があるかないかの４つの信号と、エージェントの８近傍のマスのそれぞれに目標物があるかどうかの８個の信号，さらに、直前の行動が５つの行動のそれぞれであるかどうかの５つの信号のいずれも０か１の計 17 個の信号とし、それをリカレントネットに入力した。ただし、各試行の最初での直前の行動に関する信号はすべて０とした。行動選択は、ε-greedy とし、ε の値は、学習開始時は 0.1 とし、そこから直線的に下げ、学習終了時

には 0.0、つまり、完全に greedy になるようにした。シミュレーションは、３層の場合と４層の場合について行ったが、ここでは、うまくいった４層の場合の結果について主に述べる。各層のニューロン数は、入力層（フィードバック入力を除く）17 個、下位中間層 20 個、上位中間層 10 個、出力層 5 個とし、上位中間層ニューロンの出力は, 外部からの 17 個の信号とともに次の時刻の入力信号とした。学習は、式(1)の Sarsa に基づいて生成された教師信号を用いて、BPTT にて学習を行った。ただし、学習のために過去にさかのぼるステップ数は 10 ステップで打ち切った。ニューラルネットの重み値の初期値は、上位中間層から出力層はすべて 0.0、それ以外は、-1.0 から 1,0 の乱数とした。割引率 γ は 0.9、BPTT の学習係数は 0.2 とした。

目標物は、エージェントが３回行動による状態遷移を行った後に、エージェントが存在しない３つのマスのどれかにランダムに現れることとした。この場合、たとえば、マス(0,0)にエージェントが置かれた場合、自分が迷路(a)(b)(e)のどれにいるのかの区別をすることができない。したがって、その場から動かないと、対角のマス(1,1)に目標物が置かれた場合、右に行くべきか上にいくべきかを知ることができない。しかし、もし事前に動作をし、迷路の形状を把握していれば、対角の位置に目標物が置かれても　進むべき正しい方向を選択することができる。また、目標物が置かれる直前に、内壁のあるマスにいると、壁の向こう側のマスに目標物が置かれた場合、到達するまでに３ステップかかることになり、得策ではない。したがって、目標物が出現する際には、内壁のない部屋にいる必要があるが、内壁がない場合は、逆に、何らかの形で事前に行った探索で得られた迷路の形状に関する情報を保持しておかないといけないという難しさがある。

100000 回の学習時の最後の 1000 回の平均ステップ数を見ると、３回のシミュレーションを平均して、４層の Elman ネットで 1.35、フィードバックのない４層の階層型ネットでは 1.57、3 層の Elman ネットでは 1.55 であった。学習が理想的に進み、かつ、行動が完全に greedy の場合には、隣接する部屋に目標物が置かれた場合１ステップ、対角の部屋に置かれた場合は２ステップで目標物に到達できる。目標物が隣に置かれる確率は、対角に置かれる確率の２倍であるから、目標物出現後の平均到達ステップ数が 1.33 程度になることが期待される。したがって、４層の Elman ネットの場合は、ほぼ最適ステップ数となっていることがわかる。

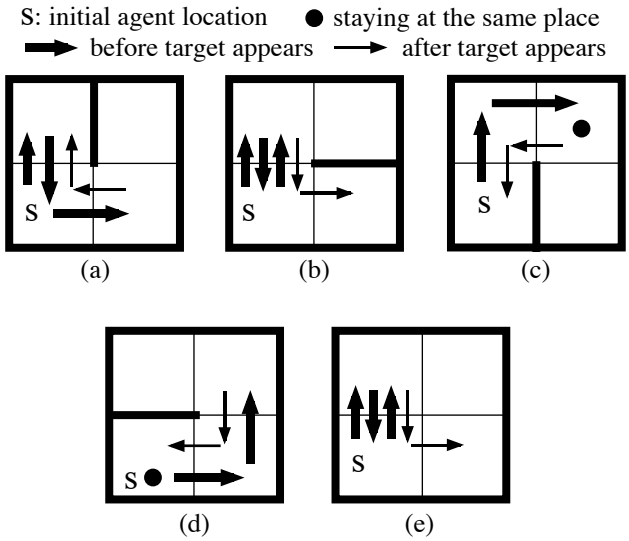

図５　学習後のエージェントの探索行動

また、４層の Elman ネットの場合の学習後の行動例を図５に示す。この図は、エージェントの初期位置を(0,0)とし、太い矢印(点)は目標物が出現する前の行動を表し、その後エージェントの対角に目標物が出現したときの経路を細線で示している。いずれの場合も、エージェントは目標物出現前に何らかの行動を行った後、出現直前には内壁のないマスにおり、対角に目標物が現れても、壁にぶつかることがない方の経路を間違うことなく取っていることがわかる。また、(a)と(b)の場合を比較すると、いずれも最初は(0,1)へ行って、再び(0,0)に戻ってくるが、(0,0)での入力は両者で同じであるにも関わらず、３回目の行動は、(a)の場合は右、(b)の場合は上と、リカレントネットが過去の記憶を保持して、(0,1)での観測値の違いを反映させた行動を実現することで、両者ともに、壁がないマスへと移動していることがわかる。このように、目標物が現れる前に、事前に知識を貯えるという行動が、単に強化学習を行うことによって獲得できることを示した。

最後に、迷路の形状を知ること自体で状態の評価値が増大し、探索行動を発現させる原動力になっているかどうかを調べるため、迷路の形状を少し複雑にした。ここでは、図６のような 3x3 の迷路を毎回ランダムに決定し、その中心にエージェントを置く。また、ゴールは試行の最初から、エージェントが存在しない場所にランダムに存在するが、エージェントからは見えないとした。観測値は、四方の壁の有無と直前の行動が

上下左右の４つの行動のそれぞれであるかどうかの計８個のみとした。Elman ネットは３層で、中間層ニューロン 30 個、初期重み値は、中間層-出力層間は 0.0、それ以外は-0.5 から 0.5 の乱数とした。BPTT の学習係数 0.01、さかのぼる最大ステップ数は 30、ε-greedy の ε は 0.1 で固定、割引率 γ は 0.9 とした。

学習結果を図６に示す。毎回、迷路の形状がランダムに選ばれるにもかかわらず、一部の迷路を除き、図のように、ゴールに到達するまで迷路をくまなく探索する行動が獲得された。このとき、迷路上のどこかに必ずゴールがあるため、探索が進んで、訪れていない場所の数が少なくなれば、次のステップでゴールに到達する可能性が大きくなる。したがって、リカレントネットがそれを学習し、各状態での最大のQ値が徐々に増加するのではないかと考えた。しかし、実際には探索を進めることによるQ値の増加は見られなかった。

そこで、ゴールの存在可能位置を四隅のみに限定すると、探索行動の学習はできなくなった。これは、たとえば、図６の(a)の場合、左上に行って戻って来たエージェントは、右に行けばゴールする可能性があるため右に行くことを学習できるが、ゴールが四隅にしかない場合は、中央から右に行っても左に行ってもゴールする可能性はなく、移動後の状態の評価値にも差がないため、左に行ったから次に右に行ってみるという行動を学習することができないと考えられる。これらは、リカレントネットがどこに行ったかを正しく覚えていれば正しい状態の評価ができると考えられること、また、リカレントネットは、簡単なカウンタや状態遷移などですら学習することが困難であるが、予め理想的な重み値をセットすると実現できる場合もあることから、リカレントネットの学習則の問題が大きいのではないかと考えられる。しかし、今後、より細かい解析をして原因を正確に追及していく必要がある。

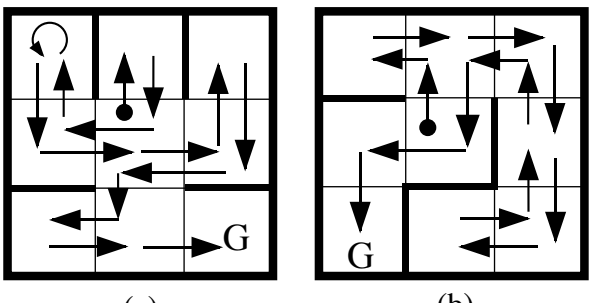

図６　ゴールが見えない場合の 3x3 の迷路の探索行動

おわりに

強化学習における「探索」を、単なる確率的な行動選択として捉えるのではなく、行動の一種と捉えることにより、逆に、強化学習によって効率的な「探索」を獲得することを提案した。そして、ゴールが不確定な問題、および、ゴールが現れる前に「探索する」必要がある問題に適用し、学習によってある程度効率的な「探索」ができるようになったことを示した。

このような考え方は、時間的および空間的に情報をいかに「抽象化」するかという問題と深く関係しているとともに、報酬がなくても行動を創発させる可能性を持つ「好奇心」というものとも密接に関係していると考えられる。これらを含めて統一的に説明できるシステムの構築が今後求められる。

謝辞

本研究は、日本学術振興会科学研究費補助金基盤研究(B)#14350227 と#15300064 および文科省海外先進教育研究支援プロジェクトの補助の下に行われた。また、カナダ、アルバータ大学 R.S.Sutton 教授には有用なコメントを頂いた。ここに謝意を表する。

参考文献

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

[2] 石井信, “強化学習におけるランダムさの自己調節”, 第３回神経情報科学サマースクール(NISS2001)テキスト(2001)

[3] 柴田克成, “強化学習とロボットの知能 -あめとむちで知能は作れるか？-“, 第16回人工知能学会全国大会論文集, パネルディスカッション「強化学習とその諸相」パネリスト原稿, 2A1-05 (2002)

[4] Sutton, R.S., Precup, D., and Singh, S., “Between MDPs and semi-MDPs: A Framework for Temporal Abstraction in Reinforcement Learning”. *Artificial Intelligence 112*:181-211 (1999).

[5] Shibata, K. and Sugisaka, M., “Dynamics of a Recurrent Neural Network Acquired through the Learning of a Context-based Attention Task”, *Artificial Life and Robotics*, Vol. 7, pp. 145-150, (2004)

[6] Rumelhart, D.E, Hinton, G.E., and Williams, R.J., “Learning Internal Representations by Error Propagation”. *Parallel Distributed Processing,* The MIT Press, pp. 318-362 (1986).