平成２０年度

修 士 学 位 論 文

題　　目　環境認識能力の変化が学習効果に及ぼす影響について

提 出 者　　室蘭工業大学大学院工学研究科

情報工学　専攻

平成　19 年　4 月　入学

氏　名　尾上　由希子

提出年月日　　平成 21 年　1 月 30日

室蘭工業大学大学院

目次

第１章　序論

1.1　本研究の背景

近年，技術の進歩に伴い多種多様なロボットが開発されつつある．ロボットは工場等で稼働する産業用のみならず，家庭や病院など日常生活環境への普及を目指したパートナーロボットや受付ロボットなど多くの高性能ロボットが開発されている[1]-[4].

ロボットの日常生活環境への普及にあたって，ロボットが用いられる環境への適応が課題の一つとして挙げられる．我々人間の生活するような日常生活環境は，乱雑で変化に富み，直面する環境は一様ではない．このような環境は動的環境と呼ばれ，ロボットがそれまでに用いられてきた工場や研究室など変化の少ない環境と異なり，時々刻々と変化し続ける．そのため，ロボットが全く同じ環境に直面することはほとんど無い．そのため，複雑で予測をすることが困難な環境である．このような動的環境においても，ロボットには自身のタスクを遂行することが期待されるため，ロボットがいかにその環境に適した行動を取るかが問題となってくる．

この問題を解決するためのアプローチの一つとして，ロボットの直面する環境を予測し行動を完全に設計するという方法が考えられる．これはロボットの設計者が，ロボットが直面するであろうあらゆる環境を事前に想定し，各環境に適した動作をロボットに設計するという方法である．この方法は，環境変化の少ない静的な環境においては，ロボットの直面する環境が限られているため有効な方法であると考えられる．しかし動的環境においては，この手法の有効性は低くなることが予測される．動的環境は時間と共に移り変わるものであり，多様で複雑なものである．そのため，設計者が全ての環境を予測しうることは困難であり，ロボットの行動設計を行う段階で設計者にかかる負荷が大きくなることが考えられる．また，仮に予測し行動を設計できたとしても，環境を完璧に予測することが困難であるため，不完全な予測によって設計に不足が生じる可能性がある．このため，動的環境においてロボットが環境に適応するためには，ロボットが自ら環境に応じた行動を獲得する事が望ましいと考えられる．

これを実現する手法の一つとして，ロボットに学習を行わせるという方法がある．この手法はロボットの行動獲得において有効な手段と考えられ，ロボットの学習に関する研究は現在に至るまで活発に行われてきている[5]-[8].

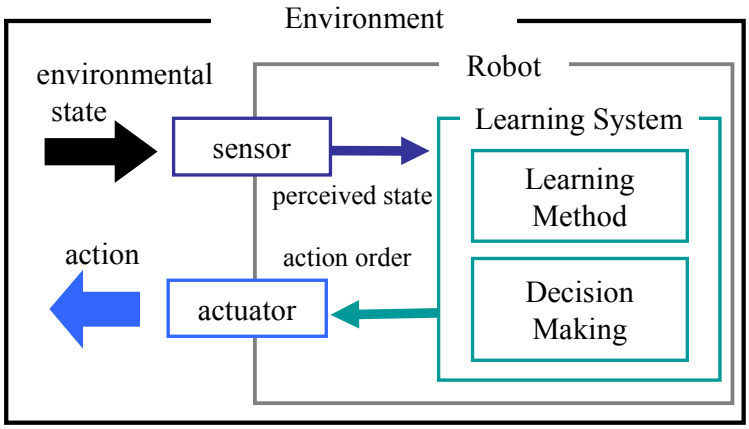


図 1.1　ロボットにおける行動学習の概要

ロボットにおける行動学習の概要を図 1.1 に示す．ロボットの行動学習において，ロボットは自身のセンサを用いて直面している環境を認識する．その認識結果と現在までの学習結果を照合し，取るべき行動について意思決定を行う．意思決定に基づいた行動の指令がロボットのアクチュエータに送られ，ロボットは行動を行う．その行動結果を受け，ロボットは学習を行う．このようなロボットの行動学習においては，学習手法に加え，センサ，及びアクチュエータの出力先であるロボットの身体構造の設定が重要となってくる．

ロボットの行動学習において用いられる学習手法として，主に機械学習[9]の手法を用いた研究が挙げられる．機械学習に含まれる学習手法では，大別して 3 種類の学習手法に分けられる[10][11].

(a) 教師あり学習
(b) 教師なし学習
(c) 強化学習

(a)の教師あり学習では，入力に対する正しい出力が教師信号として与えられ，実際に出した出力との誤差がゼロになるように学習が行われる．これは，パターン識別などの分野で多く研究され，ニューラルネットワーク[12]などに代表される．(b)の教師なし学習は，目標や正解とすべき情報が入力されない場合の学習法である．教師なし学習においては，入力信号の統計的な分布に着目し，データのクラス分けや成分への分解という形で学習が行われる．これはクラスタリングなどに代表される学習手法である．(c)の強化学習は，教師あり学習のような具体的な教師信号の与えられない学習手法である．教師信号の変わりに，実際に出力してみた結果に対して，それの評価を示す報酬が与えられる．強化学習ではこの報酬が最大になるよう学習が行われる．これらの各学習手法は，目的に応じて別々に，

あるいは併用して用いられる．

また，ロボットにおいて用いられるセンサとしては，センシングする対象とによって主に以下の 3 種類に分類される[10].

(a) 外部と内部の関係についてセンシングするもの
(b) ロボットの外部をセンシングするもの
(c) ロボットの内部をセンシングするもの

(a)のロボットの外部と内部の関係についてセンシングするセンサは，物体への距離などに代表され，ソナーセンサや GPS などが挙げられる．(b)はロボットの周りの環境などに代表され，カメラなどが挙げられる．(c)は，ロボットの内部状態を認識するものであり，ジャイロセンサやトルクセンサなどに代表される．

ロボットの身体構造については，現在様々なロボットがあるため分類することは難しい．移動ロボットのみに着目しても車輪型や二足歩行，多足歩行，蛇型など多種多様なロボットが存在する[14]-[18].

これらの学習手法やセンサを用い，更にロボットの形状（車輪型か，二足歩行か，蛇型か，ペット型かなど）を任意に決定し，各研究者はロボットに学習を行わせている．

1.2 本研究の目的

ロボットの学習に関する研究は多くの研究者によって行われている[19]-[21]．ロボットに学習を行わせる際，各研究者は任意に，適用する学習手法やロボットのセンサ，及び身体構造を選択しロボットに学習を行わせている．しかし，学習手法や用いるセンサ，身体構造が異なれば，ロボットが学習を達成できるか否かの結果は，同じ環境であっても異なってくると考えられる．これは学習手法やセンサなどのロボットの構造が異なると，それに従ってそのロボットの環境に対する適応性も異なったものになると考えられるためである．

このような影響が考えられるにも関わらず，それらロボットの構造の違いが学習に及ぼす影響についてはあまり考えられてきていなかった．これまでロボットの研究においては，既に存在するロボットに対して研究者が環境を設定し，どのようなセンサが必要であるか考え，どのような学習手法が適しているかを考慮し，学習実験を行ってきた．また，ロボットの学習においては，対象とするロボットが学習を行い，与えられるタスクを達成できるか否かの部分が最も重要となる．そのため，実際にロボットの構造の違いが学習効果に与える影響について問題点として挙がることが少なかったことが考えられる．そこで今回，ロボットの構造の違いが学習効果にどのような影響を与えるのかについて調べることを大

きな目標とする．

本論文においては，特にセンサに注目し，センサの違いが学習効果に及ぼす影響について考える．ロボットの行動学習において，いかに環境に適した行動を取れるかが重要なものとなる．そのため，環境を正確に認識するために，いかに環境を認識するのかということが重要であると考えられる．これらを考慮し本論文においては，ロボットが環境を認識する上で最も重要である要素としてセンサに注目した．

また，ロボットの環境認識能力に影響を与えるセンサの要素としてセンサの種類やセンサの解像度，サンプリング周期などが考えられる．今回本論文においては，センサの種類に注目し，センサの種類の違いによって生じるロボットの環境認識能力の違いが学習効果にどのような影響を与えるかについて調べることを目的とする．

1.3 本論文の構成

以下に，本論文の構成を述べる．

第 2 章では，本論文で用いる環境及びセンサについて定義及び説明を行うと共に，本論文で問題とするロボットの環境認識能力とセンサの関係性について概要を述べる．

第 3 章では，本論文において用いる強化学習に関して，概念及び手法について説明する．

第 4 章では，本章において述べた目的，環境認識能力の違いが学習効果に及ぼす影響に関して，その検証方法について述べる．

第 5 章では，第 4 章で述べた検証方法に基づいて，定常環境，非定常環境と環境を変え実験を行う．また，その結果及び考察を記述する．

第 6 章では，本論文全体に関する考察を記述すると共に，本論文全体を概観し，まとめを行う．併せて，本論文では扱えなかった，将来の研究課題に関して述べる．

第 2 章　ロボットのセンサと環境認識能力

2.1　本論文で用いる環境の定義

本節では，本論文で用いる環境について定義を行う．

本論文で用いる環境は，ロボットのセンサの種類の違いによる学習効果への影響を確認することを目的として用いる．具体的な環境についてではなく，一般的な環境における学習効果についての影響を確認するため，本論文で用いる環境は一般化・抽象化したものとしたものとする．そのため，一般的に用いられる言葉としての環境と異なる部分がある．本論文で用いる環境について，以下に定義を示す．また，以下において，ロボットは学習者を意味するエージェントという単語を用いて表す．

(1) 環境は，エージェント外部の全ての要素から構成される．
- エージェント外部の要素の数をN種類とすると，その環境はN種類の要素 E_i (i=1~N)によって構成される環境となる．

(2) 環境を構成する要素はそれぞれ要素の値 V_i を持つ．ここで V_i は離散値とし，i は要素の番号とする．

(3) 環境は，環境を構成する要素の値の組み合わせによって，異なる状態を取る．
- N 種類の要素によって構成される環境の場合，環境は N 次元の状態空間を持つ．N=3，$V_i = \{0,1\}$ (i=1~N)の場合の状態空間を図 2.1.1 に示す．
- V_i の取りうる値の個数を Vn_i とすると，環境のとりうる状態の数 Sn は，以下の式で表される．

$$Sn = \prod_{i=1}^{N} Vn_i \tag{2.1}$$

(4) 環境の状態は，エージェントのとる行動に従って変化する．
- 時刻 t における環境の状態を s_t とすると，エージェントの行動 a_t により，環境の状態は s_{t+1} へと推移する．
- 環境の遷移はノードとリンクによって表される．

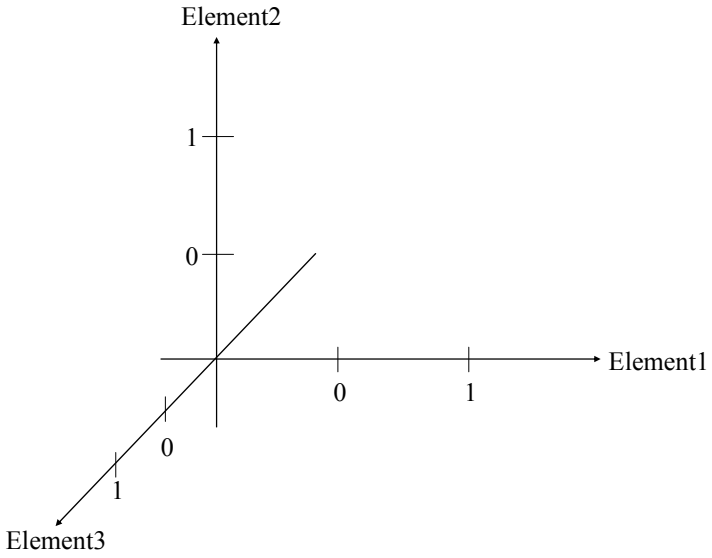


図 2.1.1　N=3，$V_i = \{0,1\}$ の場合の状態空間

本論文で用いる環境の具体的な例として，音と光という 2 種類の要素から構成される環境を示す．環境を構成する要素である音と光は，それぞれ ON（=1）と OFF（=0）の 2 値をとるものとする．この例において，環境の状態空間は図 2.1.2 で表される．

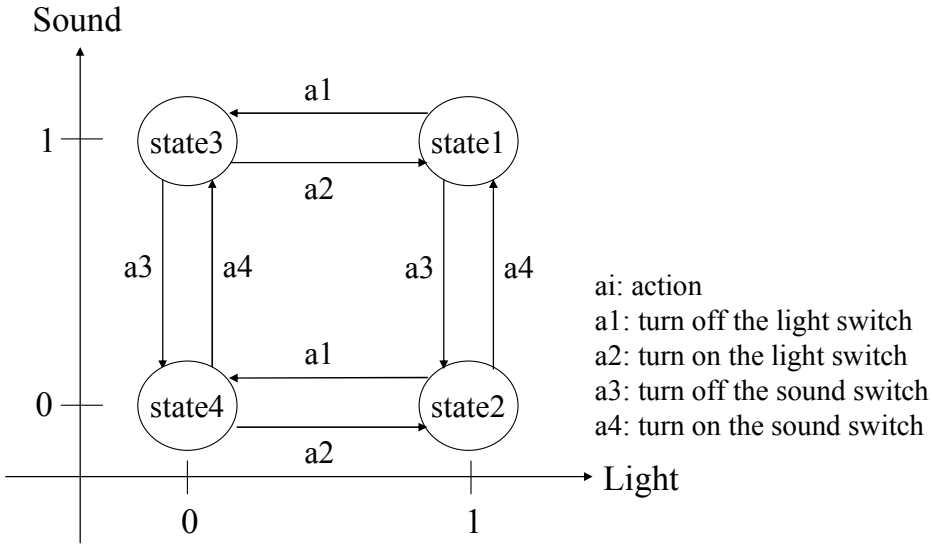


図 2.1.2　音と光における状態空間

この状態空間において，環境のとりうる状態は，光が ON で音が ON の状態（state1 とする），光が ON で音が OFF の状態（state2），光が OFF で音が ON の状態（state3），光が OFF で音が OFF の状態（state4），の 4 種類が考えられる．この環境下において s_t =state1 のとき，

エージェントが行動 a_t として a1:光のスイッチを切るという行動を取ったとする. この時, 環境の状態は s_{t+1}=state3 へと推移する.

このような環境を，エージェントが学習を行う環境として本論文では用いる.

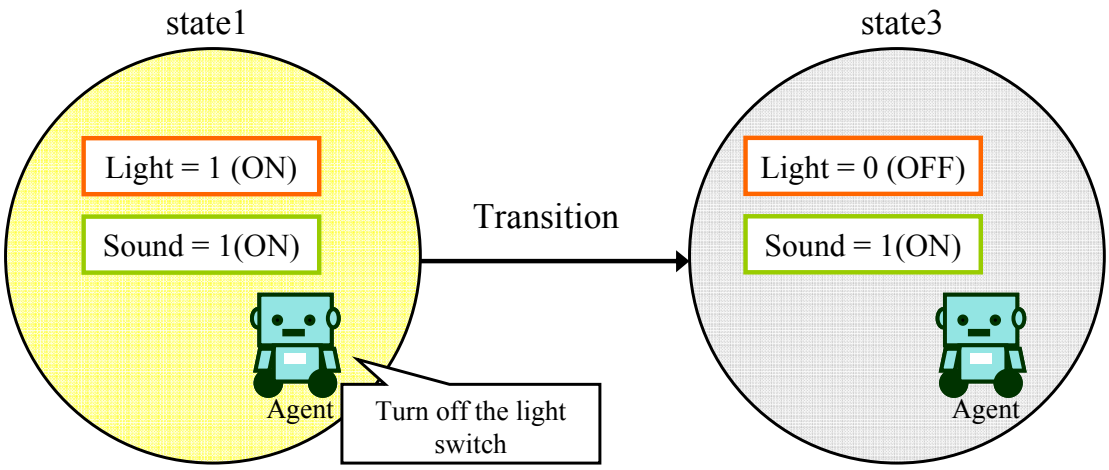


図 2.1.3　環境の状態の遷移

2.2　本論文で用いるセンサの定義

本節では，本論文で用いるセンサについて定義を行う.

本論文ではロボットのセンサの種類数の違いが学習効果に及ぼす影響を調べることを目的としている．個々の具体的なセンサによる影響ではなく，一般的なセンサについて調べることを目的としているため，センサについて一般化・抽象化をここで行う．以下に本論文で対象とするセンサの定義を行う.

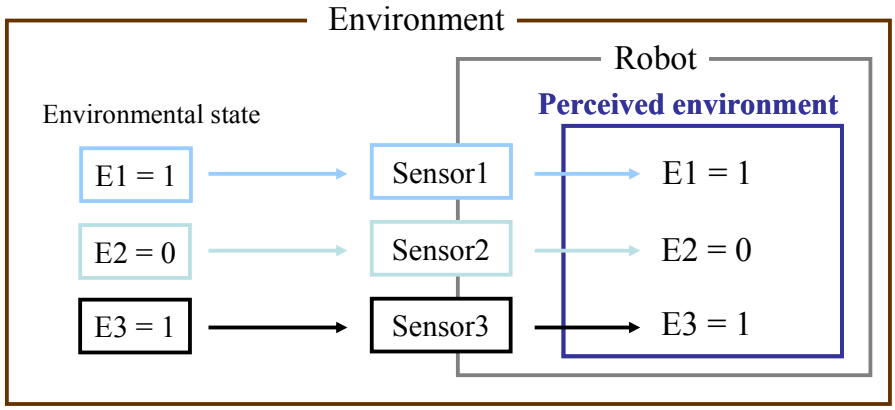
(1) 本研究で用いるセンサは，環境を構成する要素 1 種類に対し，その要素を認識することのできるセンサが 1 種類と, 1 対 1 で対応している. 環境を構成する要素 E_i に対し，それを認識するためには， E_i に対応するセンサ $Sensor_i$ が必要となる. このため，N 種類の要素によって構成される環境を完全に認識するためには，N 種類のセンサが必要となる.

(2) 学習を行うロボットは，自身の持つセンサを通して要素の値を認識する．環境を構成する要素 E_i の値 V_i は， $Sensor_i$ を通すことにより認識できる.

(3) センサが複数ある場合，各センサのセンシングする要素の重みは，全て等しいものとする．人間の場合は視覚による情報が最も重要とされるが，本論文において

は，情報における重要度は等しいものとする．

以下に例を示す．環境として E1，E2，E3 の 3 種類の要素で構成され，$V_i = \{0,1\}$ (i=1~3) をとるものを考える．この環境において，ロボットが各要素に対応する 3 種類のセンサ，Sensor1，Sensor2，Sensor3 を持つとする．ここで，Sensor1 は要素 E1 の状態を認識することができ，Senosr2 は要素 E2 を認識，Sensor3 は要素 E3 を認識することができるものとする．ロボットは，各センサを通して各要素の状態を認識し，それらの組み合わせとして環境の状態を認識することができる．この例においては要素を E1，E2，E3 としたが，これらの要素は現実世界においては音や光などに対応するものである．音と光に対応した場合，要素の値はそれぞれ F[Hz]及び C[cd]となり，対応するセンサは音センサ及び光センサとなる．



∗ E1, E2, E3 : element that compose of an environment (ex. sound, light, etc…)

図 2.2.1　センサを用いた環境認識

2.3　ロボットのセンサと環境認識

本節では，ロボットあるいは我々が認識する環境と，物理的に存在する実環境との関係について考察を行う．また，環境への適応を目指すロボットや，環境に適応するよう意識的・無意識的に学習を行っている人間について，本節では学習者を意味するエージェントという単語を用いて表す．

ロボットは，自身のセンサを通して環境の状態を認識する．人間の場合は，視覚や聴覚などに代表される五感を用いて環境を認識している．このように，実環境の中に存在するエージェントの認識する環境は，自身のセンサを通して知覚・構成した環境であり，自身

の持つセンサの種類や性能に大きく依存したものである．しかし，物理的に存在する実環境は，そこに存在するエージェントの持つセンサとは無関係に存在する．

そのため，実環境の中にはエージェントの持つセンサでは感知できず，そのエージェントには認識することの出来ない要素が存在している可能性がある．また，要素自体はエージェントの認識可能なものであっても，エージェントのセンサの分解能などの性能不足により，その要素を正しく認識できない場合も考えられる．前者の例として光センサを持たないロボットが環境の明るさについて知覚することができないということが挙げられる．また後者の例として，人間は聴覚によって音の認識はできるが可聴域を超えた音は認識できないということが挙げられる．このように，一般に物理的に存在する実環境と，エージェントの知覚する環境は異なったものであることが，エージェントの環境認識の特徴として挙げられる．

ここで，センサと環境認識についての例を挙げる．要素 E1，E2，E3 の 3 種類の要素で構成され，各要素の取りうる値が $V_i = \{0,1\}$ (i=1~3)である環境を考える．この環境において，ロボットが要素 E1 を認識することのできるセンサ Sensor1 及び E2 を認識することのできるセンサ Sensor2 を持つとする．このとき，ロボットは要素 E1 及び E2 については認識することが可能であるが，要素 E3 について認識することは不可能である．そのため，ロボットの認識する環境の状態は，実際の環境の状態とは異なるものとなる．

このような実環境と知覚環境との違いは，ロボットの持つセンサが減少し環境を構成する要素の数から離れるほど大きくなると考える．

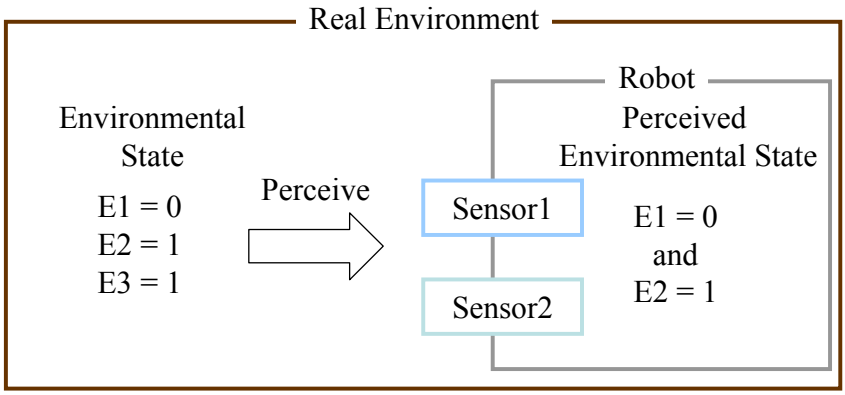


図 2.3.1　実環境とロボットの知覚する環境の違い

2.4　実環境と知覚環境の差が環境認識に与える影響

本節では，前節で示した，実環境とロボットの知覚環境との差が，ロボットにおける環境の認識においてどのような影響を与えるのかについて考察を行う．

ロボットは自身のセンサを通し実環境の認識を行う．このとき，ロボットが用いるセンサによって，実環境とロボットが知覚した環境との間に差が生じることがある．この差を生じさせるセンサの要素として，以下の 3 つについて考える．

・センサの種類
・センサの分解能
・センサのサンプリング周波数

今回，この 3 つの要素について，これらの各要素が原因となって生じるロボットの知覚環境と実環境との違いを考える．また，それがロボットの環境の認識に与える影響について考え，センサの違いが学習効果に影響を与える可能性について考察する．

2.4.1　センサの種類

まず，ロボットの持つセンサの種類の違いが環境認識に与える影響について考える．環境を構成する要素の数を N，ロボットのセンサの種類数を M とすると，環境を構成する要素の数とロボットのセンサの種類数の関係は以下の 3 つに分けられる．

(a) N = M の場合
(b) N < M の場合
(c) N > M の場合

まず(a)について考える．N=M の場合，要素 E_i (i=1~N)に対しそれを認識するのに必要なセンサ $Sensor_i$ がロボットの M 種類のセンサの中に含まれているという関係が成り立てば，ロボットは環境の全ての要素について認識することが可能である．このとき，要素 E_i (i=1~N)に対しそれを認識するのに必要なセンサ $Sensor_i$ がロボットの M 種類のセンサの中に含まれていない場合，ロボットは自身の置かれる環境の要素について認識することが不可能となるため，この場合は(c)の N>M の場合の関係と同様になると考えられる．

(b)について考えると，N<M の場合，(a)と同様に要素 E_i (i=1~N)に対しそれを認識するのに必要なセンサ $Sensor_i$ がロボットの M 種類のセンサの中に含まれているという関係が成

り立てば，この場合もロボットは環境の全ての要素について認識することが可能である．これは，環境の要素数 N に対しロボットのセンサの種類数 M が N 種類の要素を感知でき，更にそれ以上の要素について感知することができる状況であると考えられる．また，(a)と同様に要素 E_i (i=1~N)に対しそれを認識するのに必要なセンサ $Sensor_i$ がロボットの M 種類のセンサの中に含まれていない場合，ロボットは自身の置かれる環境の要素については認識することが不可能となるため，この場合は(c)の N>M の場合の関係と同様になると考えられる．

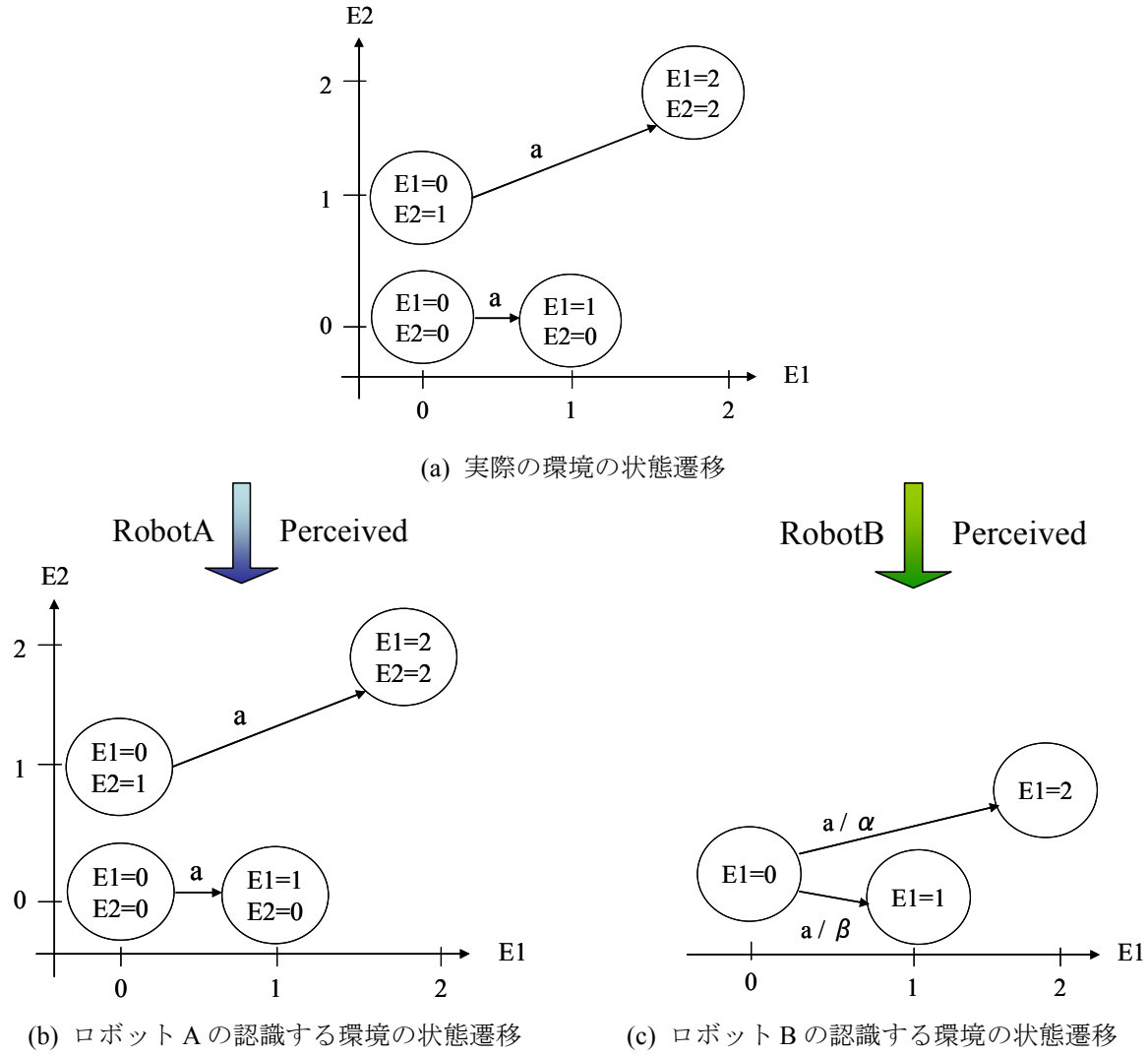
(c)について考えると，N>M の場合，ロボットは自身の持つセンサによって知覚可能な要素のみしか認識することができない．環境を構成する要素 E_i (i=1~N)について，それらの中の最大 M 種類の環境のみロボットは認識することが可能となる．これより，ロボットの知覚する環境の状態は実際の環境の一部分でしかないため，認識に不足が生じることが考えられる．

例として，2 種類の要素 E1 及び E2 で構成され，各要素の取りうる値 $V_i = \{0,1,2\}$ (i=1,2)の環境を考える．この環境において，2 種類のセンサ Sensor1 及び Sensor2 を持つロボットと，1 種類のセンサ Sensor1 のみを持つロボット B が存在したとする．ここで Sensor1 は要素 E1 を認識可能なセンサであり，Sensor2 は要素 E2 について認識可能なセンサとする．

環境の状態の遷移が，ロボットの行動 a に対して図 2.4.1(a)で示すように遷移するものであるとすると，以下のことが考えられる．

2 種類のセンサを持つロボット A は E1 及び E2 のどちらの要素の状態についても正しく認識することができる．そのため，この環境の状態を実環境のまま認識することが可能であり，実環境と知覚した環境との間に差異が生じることはないであろう．そのため，この環境において，2 種類のセンサを持つロボット A の認識する環境の遷移は図 2.4.1(a)と相違ないものであり，図 2.4.1(b)で表されるように認識されると考えられる．

しかし，1 種類のセンサしか持たないロボット B は E1 の状態しか知覚することができない．そのため，ロボット B の認識する環境は要素 E1 のみから成る環境となり，実際の環境との間に差異が生じる．図 2.4.1(a)で示す環境の遷移は，1 種類のセンサ（Sensor1）のみしか持たないロボット B には図 2.4.1(c)で示すように認識される．ここで，図 2.4.1(c)中の α 及び β は各遷移先の状態へと遷移する確率を表す．



(a) 実際の環境の状態遷移

(b) ロボット A の認識する環境の状態遷移

(c) ロボット B の認識する環境の状態遷移

図 2.4.1　センサの種類数の違いによる実環境と認識する環境の差異

各ロボットは，時刻 t に環境の状態が E1=0，E2=1 であるとき，ロボットが行動 a をとった場合に次の状態 E1=2，E2=2 に遷移する確率を，それぞれ以下のように認識する．

$$\Pr\{E1_{t+1}=2, E2_{t+1}=2 \mid E1_t=0, E2_t=1, a_t=a\}=1 \qquad (2.1)$$

$$\Pr\{E1_{t+1}=2 \mid E1_t=0, a_t=a\}=\alpha \qquad (2.2)$$

式(2.1)は 2 種類のセンサを持つロボット A の認識する，E1=0，E2=1 の状態から行動 a によって E1=2，E2=2 の状態へと遷移する確率である．また，式(2.2)は 1 種類のセンサのみを持つロボット B の認識する，E1=0 の状態から行動 a をとることによって E1=2 の状態へと

遷移する確率である．

このように，センサの種類数が実際の環境を構成する要素の数よりも少ない場合，実際は確定的である環境の遷移を確率的なものであると誤って認識することが考えられる．これは，環境の状態と行動を関連付けて覚える学習において影響を及ぼす可能性がある．

2.4.2 センサの分解能

ロボットの持つセンサの分解能の違いがロボットの環境認識に与える影響について考える．ここで，センサの分解能とはセンサによって読み取ることのできる変化の最小値とする．環境を構成する要素の取りうる値V_iの最小の変動値をδV，ロボットのセンサの分解能をRpとすると，環境を構成する要素の最小の変動値とロボットのセンサの分解能の関係は以下の 3 つに分けられる．

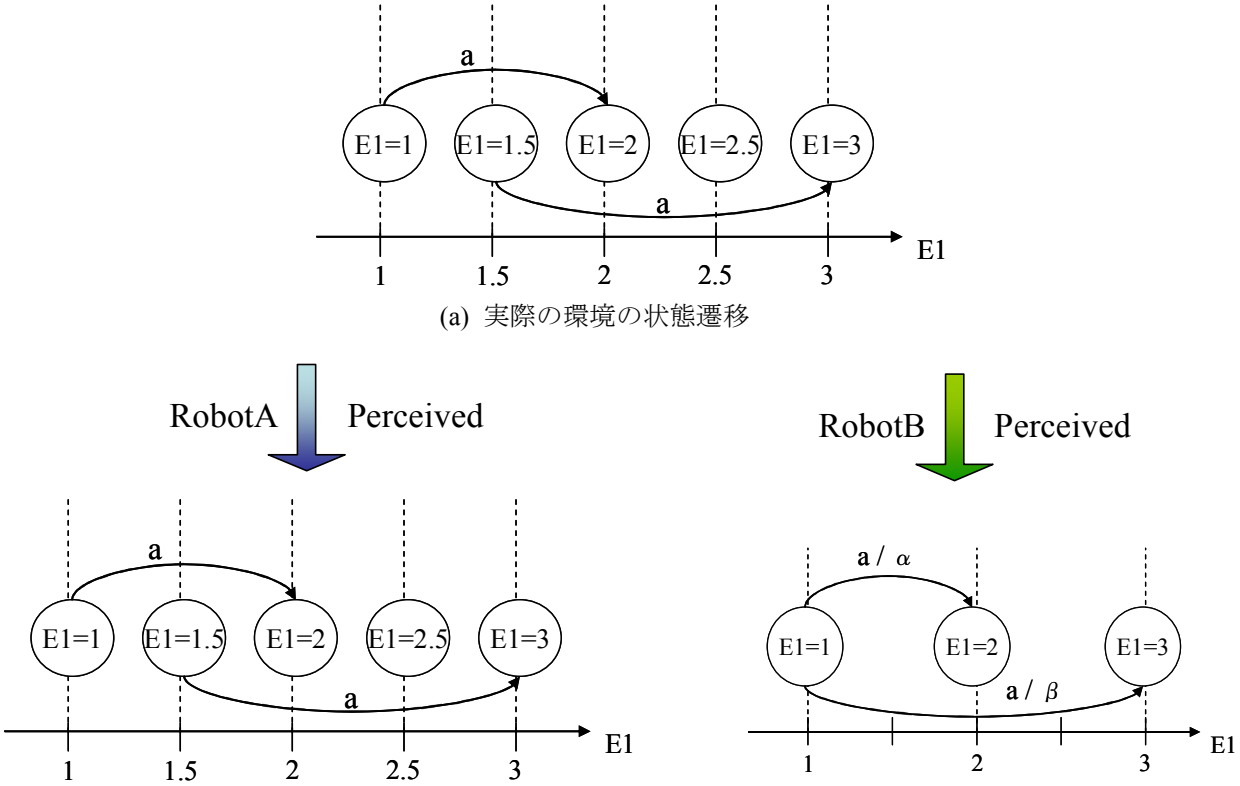
(a) $\delta V = Rp$ の場合
(b) $\delta V < Rp$ の場合
(c) $\delta V > Rp$ の場合

まず(a)について考える．$\delta V = Rp$ の場合，環境を構成する要素の変動値とセンサの分解能が等しいため，ロボットは環境の要素の全ての変化について認識することが可能であると考えられる．

(b)について考えると，この場合も(a)と同様でロボットはセンサの要素の全ての変化について認識することが可能であると考えられる．

(c)について考えると，$\delta V > Rp$ の場合，ロボットは実際の環境の要素が変動する値より粗い分解能でしか環境を構成する要素の値を認識できない．このときロボットの認識する環境は，実際の環境の要素が細かく変化した場合に追従できず，同じ状態に見えることになる．

この例を以下に挙げる．1 種類の要素 E1 で構成され，要素の値が$V_i = \{1,1.5,2,2.5,3\}$ (i=1)，すなわち$\delta V = 0.5$で 1~3 まで変動する環境を考える．この環境において，$Rp = 0.5$で E1 を認識できる分解能のセンサを持つロボット A と，$Rp = 1$で E1 を認識できる分解能のセンサを持つロボット B が存在したとする．環境の状態の遷移がロボットの行動 a に対して図 2.4.2(a)で示すように遷移するものであるとすると，次のことが考えられる．



(a) 実際の環境の状態遷移

(b) ロボット A の認識する環境の状態遷移

(c) ロボット B の認識する環境の状態遷移

図 2.4.2　センサの分解能の違いによる実環境との差異

$Rp = 0.5$ の分解能で E1 を認識できるセンサを持つロボット A は E1 の値を正しく認識することが可能である．そのため，この環境の状態を実環境のまま認識することが可能であり，実環境と知覚した環境との間に差異が生じることはない．よって図 2.4.2(a)で表される環境において，ロボット A の認識する環境の遷移は図 2.4.2(b)となり，実環境と相違ないものとなると考えられる．しかし，$Rp = 1$ の分解能で E1 を認識できるセンサを持つロボット B は E1 の状態を小数点以下抜きでしか知覚することができない．そのため，ロボット B の認識する環境は要素 E1 の細かい状態変化が省かれた環境となり，実際の環境との間に差異が生じる．図 2.4.2(a)で示す環境の遷移は，ロボット B には図 2.4.2(c)で示すように認識される．ここで，図 2.4.2(c)中の α 及び β は各遷移先の状態へと遷移する確率を表す．ロボット B の認識する環境の状態遷移が確率的なものとなっているのは，センサが $Rp = 1$ の分解能であるため E1=1 の状態と E1=1.5 の状態の区別ができないことによる．

各ロボットは，時刻 t に環境の状態が E1=1 であるとき，ロボットが行動 a をとった場合

に次の状態 E1=2 に遷移する確率を，それぞれ以下のように認識する．

$$\Pr\{E1_{t+1} = 2 \mid E1_t = 1, a_t = a\} = 1 \qquad (2.3)$$
$$\Pr\{E1_{t+1} = 2 \mid E1_t = 1, a_t = a\} = \alpha \qquad (2.4)$$

式(2.3)は，ロボット A の認識する E1=1 の状態から行動 a によって E1=2 の状態へと遷移する確率である．図 2.4.2(a)で示す環境の状態遷移において，E1=1 の状態で行動 a をとると必ず E1=2 に遷移している．そのため，ロボット A の認識する遷移確率は 1 となっている．また，式(2.4)はロボット B の認識する，E1=1 の状態から行動 a をとることによって E1=2 の状態へと遷移する確率である．

このように，センサの分解能が実際の環境を構成する要素のとりうる状態よりも低い場合，実際は細かな変化をしながら確定的に遷移する環境の遷移を確率的なものであると誤って認識することが考えられる．これは前述のセンサの種類数の場合と同様に，環境の状態と行動を関連付けて覚える学習において影響を及ぼす可能性がある．

2.4.3 センサのサンプリング周波数

ロボットの持つセンサのサンプリング周波数の違いがロボットの環境認識能力に与える影響について考える．環境の状態の変化の周波数を Fc，ロボットの持つセンサのサンプリング周波数を Fs とすると，環境の状態の変化の周波数とセンサのサンプリング周波数の関係は以下の 3 つに分けられる．

(a) $Fc = Fs$ の場合
(b) $Fc < Fs$ の場合
(c) $Fc > Fs$ の場合

まず(a)について考える．$Fc = Fs$ の場合，環境の状態の変化の周波数とセンサのサンプリング周波数が等しいため，ロボットは環境の状態変化と同じタイミングで環境の状態を認識することが可能である．そのため，ロボットは環境の状態変化を正しく認識することが可能であると考えられる．

(b)について考えると，この場合ロボットは環境の状態が変化するより速く環境の状態を認識することになる．そのため，ロボットが行動を行いその行動によって環境の状態が遷移するというプロセスにおいて，ロボットのとった行動 a に対し環境の状態が変化する前にロボットが状態を認識することが考えられる．このときロボットは行動 a を取っても環境の状態は変化しないものとして認識する可能性がある．そのため，ロボットは環境の状態変

化について正しく認識することができない可能性があると考えられる．しかし，現実の環境においては，環境の状態は連続的に変化している．そのため，$Fc < Fs$ の関係の成り立つサンプリング周波数のセンサは現実には存在しないと言える．

(c)について考えると，この場合ロボットは自身のセンサのサンプリング周波数で捉えられるタイミングでしか環境の状態を認識できない．そのため実際の環境の変化を認識できず，実際とは異なる変化として環境の状態の変化を認識する可能性がある．

この例を以下に挙げる．今回，$Fc < Fs$ の場合は現実的に問題となる場合が少ないことから，問題として扱わないものとした．1 種類の要素 E1 で構成され要素の値が $V_i = \{1,2,3\}$ (i=1) を取り，周波数 Fc =3.334，すなわち 0.3[sec]の周期で状態遷移する環境を考える．この環境において，Fs =3.334, 0.3[sec]の周期で E1 を認識するセンサを持つロボット A と，Fs =1.0, 1.0[sec]の周期で E1 を認識するセンサを持つロボット B が存在したとする．環境の状態の遷移がロボットの行動 a に対して図 2.4.4(a)で示すように遷移するものであり，時系列的に図 2.4.3(a)で変化しているものとする．これを各ロボットが認識した場合，次のことが考えられる．

0.3[sec]周期で E1 を認識するセンサを持つロボット A は，環境の状態について，実際の変化と同様の 0.3 秒間隔で認識することができる．そのため，この環境の状態変化をそのまま認識することが可能であり，実環境と知覚した環境の間に差異が生じることはない．そのため，ロボット A が認識する，環境の状態の時系列的な変化は図 2.4.3(b)で示すものとなり，行動 a についての遷移は図 2.4.4(b)であると認識することができる．しかし，1[sec]周期で E1 を認識するロボット B は，E1 の 0.3 秒間隔での変化を認識することができない．そのため，ロボット B の認識する環境は，実際の環境状態と異なった変化をし，見落としの多いものとなると考えられる．ロボット B が時系列的に認識する環境の状態の変化は図 2.4.3(c)で示すものとなる．この例において，環境の状態は，t=0.3n（n : 自然数）の時刻 t[sec]において E1=V_t のとき，周期である 0.3[sec]の間 E1=V_t を保ち続けるものとしている．また，ロボット B は環境の状態の遷移について図 2.4.4(c)のように行われるものと認識する．ここで，図 2.4.4(c)中の α 及び β は各遷移先の状態へと遷移する確率を表す．

各ロボットは，時刻 t において環境の状態が E1=3 であるとき，ロボットが行動 a をとった場合に次の状態 E1=1 に遷移する確率を，それぞれ以下のように認識する．

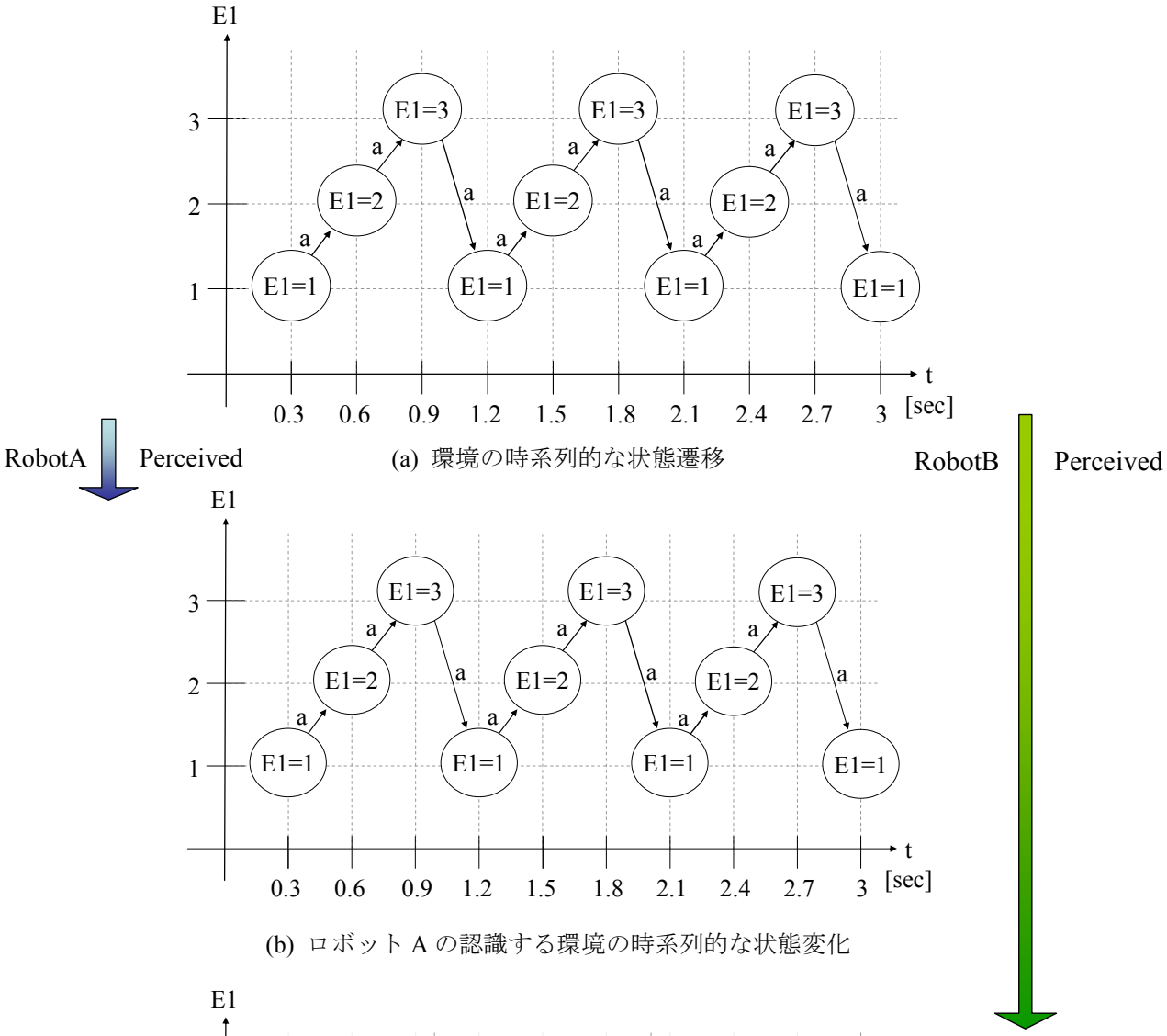
$$\Pr\{E1_{t+1} = 1 \mid E1_t = 3, a_t = a\} = 1 \qquad (2.5)$$

$$\Pr\{E1_{t+1} = 1 \mid E1_t = 3, a_t = a\} = \alpha \qquad (2.6)$$

式(2.5)は，ロボット A の認識する E1=3 の状態から行動 a によって E1=1 の状態へと遷移する確率である．図 2.4.3(b)で示す環境の状態遷移において，E1=3 の状態で行動 a をとると必ず E1=1 に遷移している．そのため，ロボット A の認識する遷移確率は 1 となっている．また，式(2.6)はロボット B の認識する，E1=3 の状態から行動 a をとることによって E1=1

の状態へと遷移する確率である．

このように，センサのサンプリング周波数が実際の環境の状態遷移の周波数よりも低い場合，環境の状態変化について見落としが多く，実際の遷移が認識できないことが考えられる．これより，本来因果関係の分かる状態遷移について，確率的なものと誤って認識する可能性がある．この場合についても環境の状態と行動を関連付けて覚える学習において影響を及ぼす可能性がある．

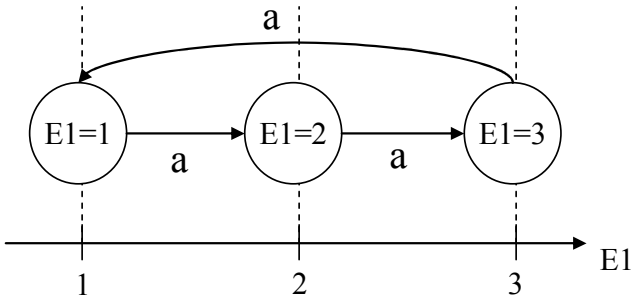


(a) 環境の時系列的な状態遷移

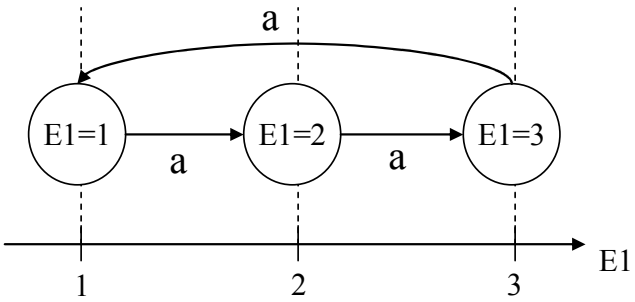
(b) ロボット A の認識する環境の時系列的な状態変化

(c) ロボット B の認識する環境の時系列的な状態変化

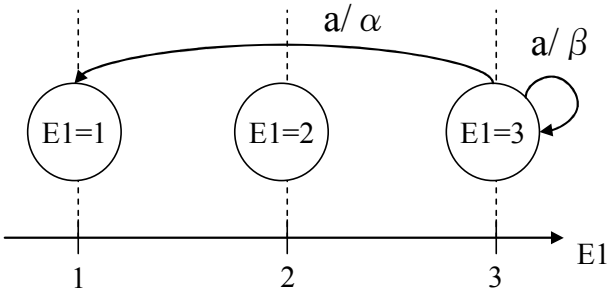
図 4.2.3　センサのサンプリング周波数の違いによる実環境との時系列的な差異



(a) 行動 a に対する環境の状態遷移



(b) 時系列を通してロボット A の認識する，行動 a に対する環境の状態遷移



(c) 時系列を通してロボット B の認識する，行動 a に対する環境の状態遷移

図 2.4.4　サンプリング周波数の違いによる実際の環境の状態遷移との差異

2.5 まとめ

本章では，ロボットのセンサと環境認識能力の関係について述べた．まず，本論文で用いる環境とロボットのセンサについての定義について説明し，ロボットによる知覚環境と実環境との違いについて考察した．そしてその知覚環境と実環境の違いによって環境認識がどのように変わるかについて考え，ロボットが学習を行う際にそれらが影響する可能性について述べた．

第 3 章　強化学習

本章では，本論文における実験で用いる学習手法，強化学習[22]について述べる．

3.1　概要

強化学習は，ある環境において学習を行う者が，現在の状態を観測し，取るべき行動を決定する問題を扱う，機械学習の一種である．学習者は環境との相互作用を繰り返しながら，行動の結果によって環境から与えられる報酬をもとに自らの行動を改善する．強化学習はこのように試行錯誤的に行われる学習法であり，学習者は受け取る報酬を最大化することを目標に学習を行う．

強化学習が他の機械学習と大きく異なる点は，

(1) 学習に際して，正解が与えられない（教師あり学習と異なる）

(2) 学習する内容が，学習者の行動に依存する（能動性）

にある．したがって，

(1) どういう行動が望ましいかを予め明確化する必要が無く，

(2) 学習者自身が，学習すべき内容を能動的に決定し，

学習を進めることが可能となることが最大の特長である．

強化学習において，設計者は学習者に対し「何をすべきか」という目標を報酬という形で指示しておけば，「どのように実現するか」は学習者によって自動的に獲得される枠組みとなっている．このため，ロボットの行動獲得という目的への応用が期待され，また多くのロボットの学習において用いられている手法である[23]-[26].

3.2　概念及び用語説明

3.2.1　環境とエージェントの相互作用

強化学習におけるエージェントと環境の相互作用の概念図を図 3.2.1 に示す．ここで，学習者は，学習者を意味するエージェントという単語を用いて表記する．環境から知覚される状態 s_t において，エージェントが何らかの行動 a_t を取った場合，環境からその行動の良

し悪しを数値に写像した報酬r_tを受け取り，この報酬を基にエージェントは学習を行う．また，学習者の知覚する環境の状態は，次の状態であるs_{t+1}に遷移する．

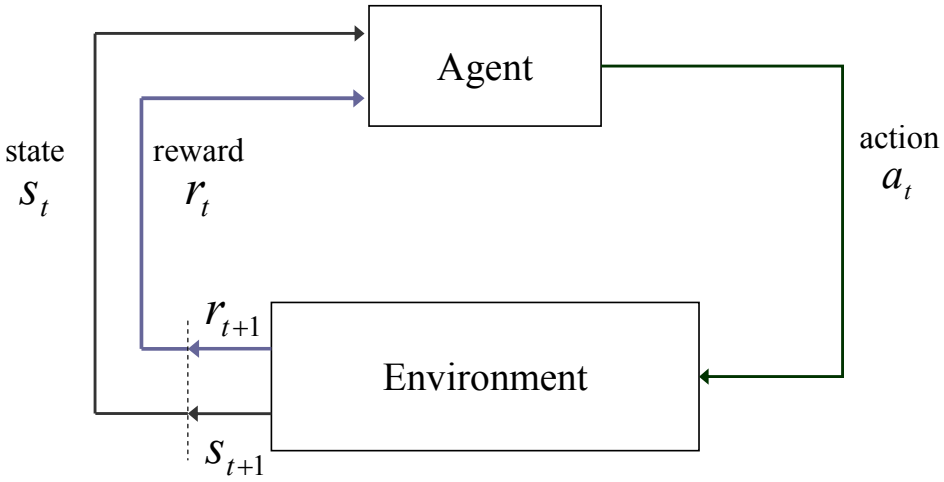


図 3.2.1　エージェントと環境の相互作用

3.2.2　強化学習の構成要素

強化学習の主な構成要素として，方策・報酬関数・価値関数・環境のモデルがある．これら構成要素及び強化学習で頻繁に用いられる単語について本節では説明を行う．

・エージェント

強化学習の枠組みにおいて，学習と意思決定を行う者を指す．ロボットの行動学習などの場合においてはロボットを指す．

・環境

エージェント外部の全てから構成される．エージェントが相互作用を男行う対象である．

・方策

行動の選択方法をあらわす．ある時点でのエージェントの振る舞い方を定義するものである．方策は，環境において知覚した状態から，その状態にあるときに取るべき行動への写像である．この方策は一般的に確率的である．

・報酬関数

報酬関数は目的を定義する．エージェントがとった行動に対する評価を数値化したものであり，報酬はその状態におけるエージェントの行動

の望ましさを表している．強化学習のエージェントの唯一の目的は受け取る報酬を最大化することである．エージェントが報酬関数を変更することはできないが，方策を変更する指針として使うことができる．報酬関数も一般的に確率的である．

・価値関数

報酬関数が即時的な行動の良さを表すのに対し，価値関数は最終的な行動の良さを指定する．状態の価値とは，エージェントがその状態を基点として将来にわたって蓄積することを期待する報酬の総量である．報酬はその環境が即時的で固有の望ましさを決定するのに対して，価値はその後に続きそうな状態群とそれらの状態群で得られそうな報酬を考慮に入れた上での長期的な望ましさを示すものである．

意思決定を行い，その決定の結果を評価するには，エージェントは即時的な報酬ではなく，長期的な望ましさを意味する価値に最も関心を払う必要がある．そのため，行動の選択は価値を判断した結果に基づいている．最終的に最大の報酬を得るためには，報酬ではなく，もっとも高い価値を持つ状態につながるような行動を見つけ出すことが必要となる．故に強化学習問題では，価値を評価・推定することが最も重要であるとされている．

3.2.3　強化学習の流れ

強化学習は，3.2.1 で示した相互作用により行われる．その流れを具体的に述べる．

強化学習の流れを図 3.2.2 に示す．環境から知覚した状態s_tによって，エージェントは自身の選択可能な行動から，行動a_tを選択し実行する．この際エージェントは状態s_tにおける行動価値に基づき，自身の適用している行動の選択手法を用いて行動選択を行う．その結果得られた報酬r_tを基に，エージェントは状態s_tにおける行動a_tの価値（行動価値）$Q(s_t, a_t)$を更新する．この際エージェントは行動評価関数を用いて価値の更新を行う．これによってエージェントは学習を行い，エージェントが次回同様の状態に直面した際においての行動選択に活かす．

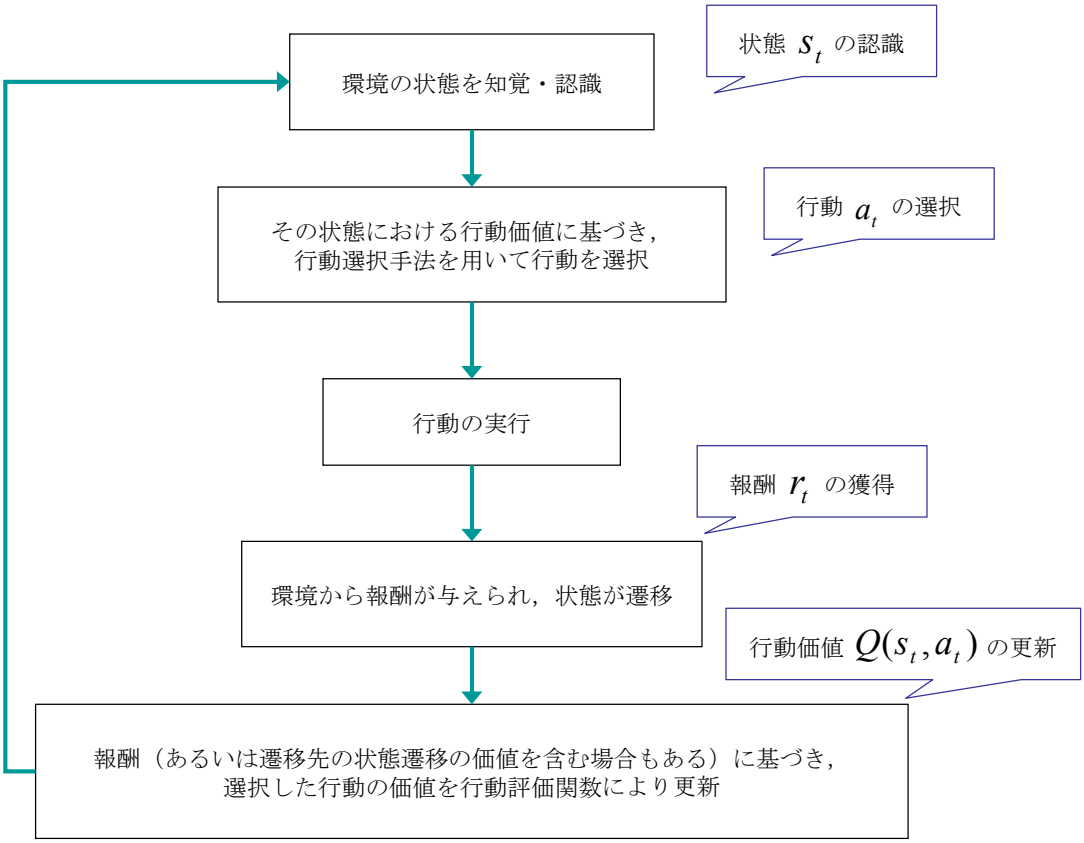


図 3.2.2　強化学習の流れ

3.3　行動選択手法と行動価値の評価方法

3.3.1　行動選択手法

行動選択手法とは，エージェントが認識した環境の状態 S において，とる行動を選択する際に用いられる手法である．

強化学習における行動選択の際に重要となるのは，単に現在の推定価値が最大となる行動を選択するのみではなく，より価値の高い行動を求める探索を行うことである．両者間のトレードオフを exploration-exploitation 問題と言う．現在までに得た知識を利用し，現在の推定価値が最大となる行動を選択することは，最終的に得られる報酬量を大きくするという目的のために重要である．しかし，より価値の高い行動を求め探索を継続することは，

局所解に陥らず方策の正しい価値推定を行うため，また，非定常環境において環境の変化に追随するために有効である．知識利用と探索のどちらか一方に偏って学習を行った場合，学習がうまく進まなくなることが予測される．そのため，このバランスをどうとるかが学習を行う上で重要となる．

探索と知識利用の両立という観点から，比較的よく用いられる行動選択手法として，ε-greedy と softmax 手法がある．以下にその行動選択手法を説明する．

・ε-greedy 法

ε-greedy 手法は，現在推定される行動価値が最も高い行動（グリーディな行動）を(1-ε)の確率で選択するか，小さい確率εで一様に任意の行動を選択するという手法である．(1-ε)の確率で行う，現在の推定価値が最も高い行動の選択が exploitation に相当し，εの確率での，ランダムな行動選択が exploration に相当する．εが小さいほど，現時点で最適とされる行動が行われる回数は多くなるが，真に最適な行動を見つけ出すまでに時間がかかってしまう．それゆえ，εの値を調整し，探索と知識利用のバランスの取り方を考える必要がある．ε-greedy 法の欠点として，確率εにおける行動選択の際に，行動価値に関わらず一様に任意に行動を選択することが挙げられる．このため，確率εでの行動選択において，ほとんど最悪と思われる行動を選択する可能性とほとんど最適行動に近い行動を選択する可能性が同じくらいに高くなるということがある．

・softmax 法

softmax 手法は，推定される行動価値に基づいた確率で行動を選択するという行動選択手法である．一般的に Gibbs 分布，あるいは Boltzmann 分布に基づいて行動が選択される．具体的には，t 回目の試行における行動aの行動価値$Q_t(a)$が与えられた場合，行動aを選択する確率$\pi(a)$は次式で与えられる．

$$\pi(a) = \frac{e^{Q_t(a)/\tau}}{\sum_{b=1}^{n} e^{Q_t(b)/\tau}} \qquad (3.1)$$

ここでτは温度と呼ばれる正定数で，温度が高い場合には，すべての行動がほぼ同程度に起こるように設定され，低い場合には、価値の推定が異なる動作の選択確率の差がより大きく異なるように設定される．

3.3.2 行動価値の評価・推定手法

強化学習における行動価値の推定手法について説明する．

行動aをとった際の平均報酬を行動aの真の価値$Q^*(a)$とし，t 回目の試行におけるその

推定量を$Q_t(a)$とする．強化学習において，学習エージェントは行動aの真の価値そのものを知ることはできず,行動によって得られる報酬からそれを推測した$Q_t(a)$を学習し行動選択に用いる．この行動価値推定の方法の１つとして，標本平均化手法がある．

標本平均化手法は，その行動が選ばれたときに実際に受け取られた報酬を平均化してゆく方法である．t 回目の試行において，それまでの間に行動aがk_a回選択されていて，各回で得られた報酬が$r_1, r_{2,} \cdots, r_{k_a}$とすれば，行動$a$の推定価値$Q_t(a)$は次式で求められる．

$$Q_t(a) = \frac{r_1 + r_2 + \cdots + r_{k_a}}{k_a} \qquad (3.2)$$

k_a=0 の場合には，$Q_t(a)$を，$Q_0(a)$=0 のようなデフォルト値に設定する．$k_a \to \infty$の極限において，大数の法則から，$Q_t(a)$は$Q^*(a)$に収束する．

本節で述べた行動選択手法と行動価値の推定手法の拡張が，様々な強化学習の手法となる．その強化学習手法の一般的な手法のうち，本研究で用いたものについて次節で述べる．

3.4　一般的な強化学習手法

本節では強化学習法においてよく用いられ，本研究においても使用した行動価値の評価手法について説明する．

3.4.1　強化比較法

強化比較法は，状態遷移の無い比較的単純な強化学習課題に用いられる手法である．強化比較法では，与えられた報酬の大小を評価するための基準レベルをリファレンス報酬と呼び，現在までに受け取った報酬の平均値を用いる事が多い．

この基準レベルより大きい報酬が得られた行動は，良い行動と判断され以後この行動をとる確率が上がる．一方，基準レベルを下回る報酬につながった行動に関しては，以後この行動をとる確率を下げることにより次第に報酬の大きな行動が選択される傾向が強まる．

実際の行動選択に当たっては，通常 softmax 手法(2.2.1 参照)が用いられる．この場合，t 回目の試行において行動aを選択する優先度$p_t(a)$を用い，行動aを選択する確率$\pi(a)$は次式で与えられる．

$$\pi_t(a) = \frac{e^{p_t(a)}}{\sum_{b=1}^{n} e^{p_t(b)}} \qquad (3.3)$$

また，行動aを選択する優先度及びリファレンス報酬$\bar{r}$は次式によって更新される．

$$\begin{aligned} p_{t+1}(a_t) &= p_t(a_t) + \beta[r_t - \bar{r}_t] \\ \bar{r}_{t+1} &= \bar{r}_t + \alpha[r_t - \bar{r}_t] \end{aligned} \qquad (3.4)$$

ここで，β は正のステップサイズ・パラメータを示し，優先度に関わる報酬の重みを表す．また α（0<α<1）はリファレンス報酬の学習率を示している．

3.4.2 追跡手法

追跡手法は，行動価値推定と行動優先度の両方を利用した学習手法である．優先度は現在の行動推定価値に従ったグリーディな行動を「追いかける」目的で使用される．t 回目の試行で行動aを選択する確率$\pi(a)$を行動優先度として用いられる事が多い．
毎回の試行の直後，グリーディな行動が選ばれる可能性がより高くなるように，この確立値は更新される．t 回目の試行の後，t+1 回目の試行に対するグリーディな行動（複数個ある場合にはその中からランダムに選んだ１つ）を$a^*_{t+1} = \arg\max_a Q_{t+1}(a)$とする．この場合，

行動$a_{t+1} = a^*_{t+1}$の選択確率は

$$\pi_{t+1}(a^*_{t+1}) = \pi_t(a^*_{t+1}) + \beta[1 - \pi_t(a^*_{t+1})] \qquad (3.5)$$

で表され，確率１に向かって β の比率で増加させられる．残りの行動の選択確率は，全ての$a \neq a^*_{t+1}$に対して，次のように 0 に向かって減少される．

$$\pi_{t+1}(a) = \pi_t(a) + \beta[0 - \pi_t(a)] \qquad (3.6)$$

行動価値$Q_{t+1}(a)$は，標本平均化手法（3.2.2 参照）などを用いて更新される．

3.4.3 Q 学習

多くの強化学習手法は，離散化された状態空間と時間の上に組み立てられている．本論文で主に用いるQ学習という学習法は,継続する状態間の効用の差分を利用することから，時間的差分学習（TD 学習）と呼ばれる強化学習手法に分類される．

時間的差分学習とは，環境のダイナミクスのモデルを用いずに経験から直接学習することができ，最終結果を待たずに他の推定値の学習結果を一部利用し，推定値を更新する学習法である．

Q 学習は，行動価値（ある状態である行動をとることの価値で，一般的に Q 値と呼ばれる）用いた，方策オフ型の TD 学習手法であり，ある方策（挙動方策と呼ばれる）に基づいて行動しながら，最適方策を学習する点に特徴がある．例えば行動選択手法として ε-greedy 手法を用いた場合，ε-greedy 手法に基づく行動決定を行いながら実際には最適方策を学習する．

１ステップ Q 学習における行動価値の推定の改善は次式によって行われる．

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha[r_{t+1} + \gamma \max_a Q(s_{t+1},a) - Q(s_t,a_t)] \qquad (3.7)$$

ここで，s_t は現在の状態，a_t は採用した行動，r_{t+1} は行動によって得られた報酬を示し，s_{t+1} は行動後の新しい状態，a は新しい状態において選択される行動である．また，$Q(s_t,a_t)$ は状態 s_t における行動 a_t の行動価値推定を示し，α(0<α<1)は学習率，γ(0≦γ<1)は割引率を表す．

3.5　まとめ

本章では，本研究の実験において用いる学習手法である強化学習について説明し，その概要と用語及び各学習手法について述べた．

第 4 章　環境認識能力が学習効果に及ぼす影響の検証方法

本章では，環境認識能力が学習効果に及ぼす影響をどのように検証し確認するか，その方法についての説明を行う．

4.1　検証方法

本節では，ロボットの環境認識能力が学習効果に及ぼす影響の判定法について述べる．

本論文においては，ロボットの環境認識能力に影響を及ぼす要素として，センサの種類数に着目し，センサの種類数の違いが学習効果にいかに影響を及ぼすかについて検証することを目的としている．その目的を達成するためのアプローチとしての検証方法を以下に示す．

今回，ロボットの環境認識能力に差を生じさせるため，複数台のロボットを用意し，各々の持つセンサの種類数を異なったものとして設定する．ここで，今回はセンサによる影響のみに着目しているため，各ロボットの用いる学習手法及び身体構造，選択できる行動は同じものを用い統一させる．また，本論文では環境認識能力に影響を与えるであろうセンサの要素のうち，センサ種類数の違いに焦点を絞っている．そのため，センサの分解能及びサンプリング周期についても各ロボットについて同様のものとし，この 2 つに関しては，センシング可能な要素から与えられる情報をそのまま認識できるものとする．

これらセンサの種類数の異なる複数台のロボットを同一の環境において学習を行わせ，その結果を比較する．この学習結果の比較により，センサの種類数の違いによって生じる環境認識能力の違いが学習効果に及ぼす影響について検証する．

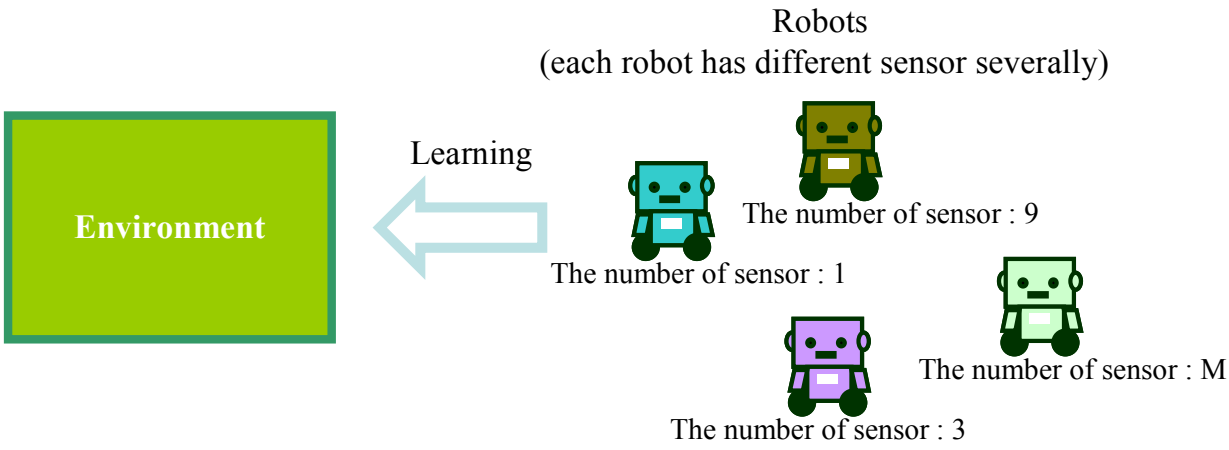


図 4.1　検証方法の概念図

4.2　検証を行う環境とロボットの設定

本節では，検証を行う際に用いる環境とロボットについての設定を述べる．

検証方法は 3.1 で述べたとおりであり，同一の環境に対してセンサの種類数の異なる複数台のロボットが学習を行うという方法を取る．ここで検証を行う環境及びロボットについて考える．

本論文ではセンサの種類数の違いによる学習効果への影響に着目している．環境の設定においてセンサの種類数の違いに関係するのは，環境を構成する要素の数である．学習を行うロボットの持つセンサの種類数 M について，最大を M_{max}，最小を M_{min} とし，環境の構成要素数を N とすると，M_{max} と N，M_{min} と N について次の関係が考えられる．

・M_{max} と N の関係

(1)　$N < M_{max}$
(2)　$N = M_{max}$
(3)　$N > M_{max}$

・M_{min} と N の関係

(1)　$N < M_{min}$
(2)　$N = M_{min}$

(3)　$N > M_{\min}$

ここで，環境を構成する要素 E_i (i=1~N)とロボットのセンサ $Sensor_j$ (j=1~M)において，環境の構成要素 E_k はロボットのセンサ $Sensor_k$ によって知覚されるものとする．

$M_{\max}$ と N の関係において，$N > M_{\max}$ の場合，学習を行う複数台のロボットにおいて，どのロボットも完全に環境を認識することが不可能となる．そのため，検証において，ロボットが完全に環境を認識できた場合の学習効果を確認することができなくなる．今回の検証は学習効果の比較を目的としているため，環境を完全に認識できる場合の学習効果が確認されることが望ましい. そのため, $M_{\max}$ と N の関係は $N < M_{\max}$ あるいは　$N = M_{\max}$ であることが検証に適していると考えられる. また，$M_{\min}$ と N の関係において，$N = M_{\min}$ あるいは $N < M_{\min}$ の場合，学習を行う複数台のロボットにおいて，すべてのロボットが完全に環境を認識することが可能となる．この場合，各ロボットの環境認識能力に差が生じないことが考えられる．これは本検証の目的と外れるため，$M_{\min}$ と N の関係は $N > M_{\min}$ であることが適していると考えられる．

よって本論文における検証では，環境の構成要素数とロボットのセンサの種類数について以下の関係が成り立つ場合が適しているものと考える．

$$M_{\min} < N \leq M_{\max} \tag{4.1}$$

また，検証を行う環境において，環境の取りうる状態の数は以下のように設定する．環境を構成する要素 E_i の値を V_i とする．この V_i の取りうる値の数を Vn_i とすると，環境の状態は要素の値の組み合わせによって表現されるため，環境のとりうる状態の数 Sn は，以下の式で表される．

$$Sn = \prod_{i=1}^{N} Vn_i \tag{4.2}$$

検証で用いる環境の状態の数は(3.2)式に従って設定する．

この設定において検証を行うことにより，センサの種類数の違いによる学習効果への影響が確認できると考える．

4.3 まとめ

本章では，ロボットの環境認識能力が学習効果に与える影響の検証について述べた．まず，検証方法について説明し，次いで検証を行う環境とロボットの設定について説明した．

以降，本論文で行う実験はこの検証方法に基づいて行われる．

第 5 章　実験

5.1　実験概要

本節では，本論文で行う実験の概要について説明する．

実験は，実機ではなくシミュレーションを用いて行い，以下の 3 種類の実験を行う．また，各実験において 4 章で述べた検証方法を用いる．

実験 1：定常環境における実験 (1)

実験 2：定常環境における実験 (2)

実験 3：非定常環境における実験

実験 1 では，学習を行う環境を，変動の無い定常環境に設定し実験を行う．

実験 2 では，実験 1 の結果より実験方法を検討し，実験 1 とは異なる設定での実験方法及び定常環境において実験を行う．

実験 3 では，学習を行う環境を，確率的に変動する非定常環境に設定し実験を行う．

5.2　定常環境における実験 (1)

本節では，エージェントが定常環境において学習を行った実験について述べる[27].

5.2.1　実験の目的

本実験では，定常環境においてセンサの種類数の違いが学習効果に与える影響を調べることを目的としている．

5.2.2　実験方法

実験方法は，4 章で述べた検証方法に基づく．N 種類の要素によって構成される環境に対して最大 M 種類（N≦M）の異なる種類のセンサを持つ複数のエージェントが学習を行うという方法を用いる．

今回の実験は次のような流れで行った．1 種類の環境において，各エージェントが R 回数学習を行った後，各エージェントの知識を保持したまま，スタート地点の状態に戻す．そして再び R 回数学習を行わせる．これを S 回繰り返した後，各エージェントの知識及び得た報酬を初期化し，学習を行う環境を異なるものに設定する．そして再びエージェントに学習を行わせる．今回，R 回数の学習の後スタート地点に戻すという方法を取ったのは，常に 1 度同じ状態に戻り学習を行うことにより，学習が行なわれ易くなると期待したためである．また，環境の種類を変えて学習を行わせるよう設定したのは，環境によってあるエージェントが学習しやすいなど傾向が出る場合が予測されるため，複数種類の環境で学習を行うことにより，それらの傾向に囚われず学習結果の比較を行うことが可能となると考えられるためである．これらの流れを図 5.2.1 に示す．ここで，全てのエージェントが 1 回学習を行う状態を 1-trial, 各エージェントが R 回学習を行いスタート地点に再び戻るまでを 1-time，1 種類の環境において S 回スタート地点に戻り，各エージェントが R・S 回の学習を完了するまでを 1-set としている．

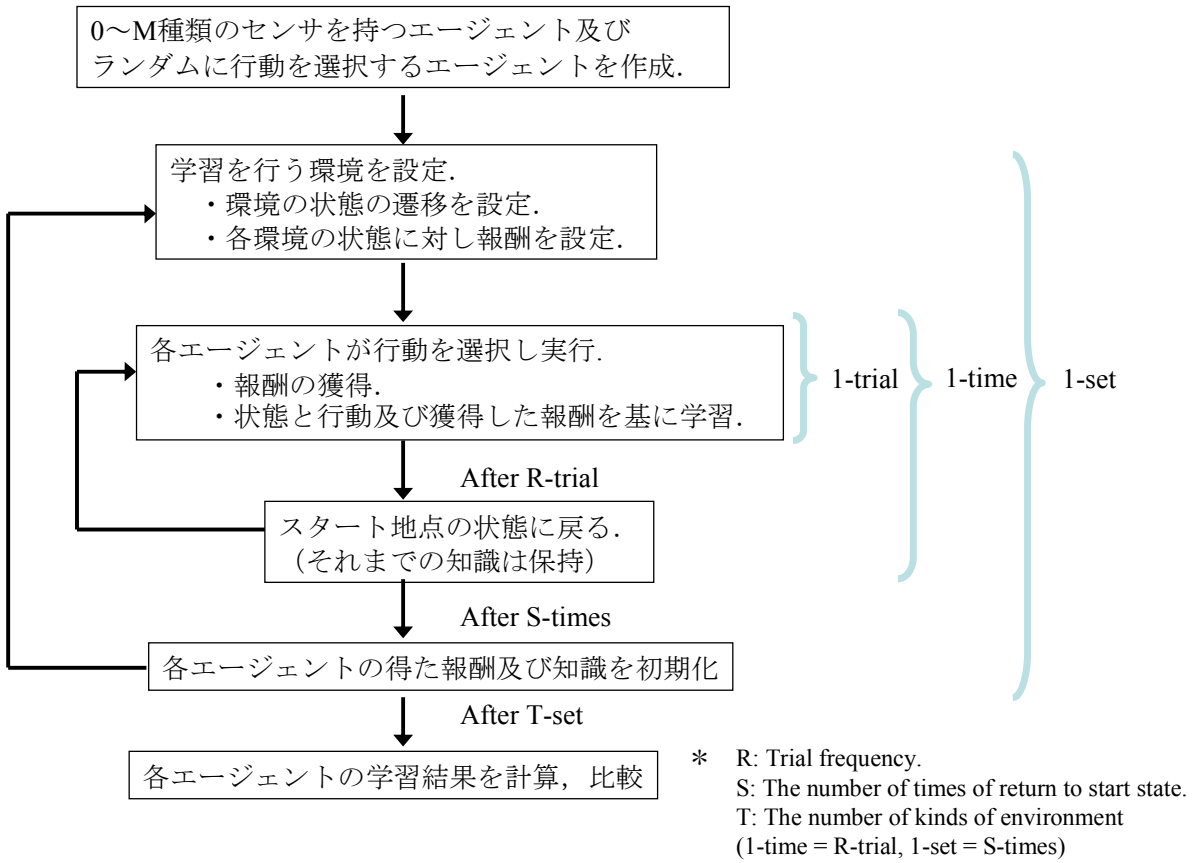


図 5.2.1　実験の流れ

5.2.3　実験に用いるタスク

本実験においてエージェントの行うタスクは，最大の報酬が得られるルートの探索とする．エージェントは環境の各状態において行動を選択することにより，次の状態へと遷移する．遷移先の状態が良いものであればエージェントは高い報酬を得，悪いものであれば低い報酬，あるいは負の報酬を得る．また，その状態のみを見ると悪いものであったとしても，更にそこから遷移すると良い状態へと行けるなど，本タスクにおいては先を見据えた学習が必要となる．各状態へ移動することにより得られる報酬の総和を有限移動回数内で最大にすることがエージェントの目的となる．

ここで，環境の各状態における報酬の設定は，次のように設定する．現実世界において高い報酬が得られる状態を考えた場合，一般的に到達することが難しく，また留まり続けることも難しい場合が多い．そのような高い報酬の得られる状態に到達し，またそこに留まるためには，何らかの知識や学習など賢さが求められる場合が多いことが考えられる．今回はそのような現実世界をモデルとし報酬の設定を行った．本実験において，環境の各状態はそれぞれ番号（STATE0，STATE1…）が与えられており，状態 i への移動に対して与えられる報酬 Rwd_i を式(5.1)に従って設定する．

$$Rwd_i = D_{si} \cdot w_{reach} + (D_{si} - D_{si}avg) \cdot w_{avg} + D_{ii} \cdot w_{keep} + (D_{ii} - D_{ii}avg) \cdot w_{avg} \qquad (5.1)$$

$$D_{si}avg = \frac{1}{N}\sum_{i=1}^{N} D_{si} \quad , \quad D_{ii}avg = \frac{1}{N}\sum_{i=1}^{N} D_{ii}$$

式(5.1)において，最初の 2 項，$D_{si} \cdot w_{reach} + (D_{si} - D_{si}avg) \cdot w_{avg}$ は i 番目の状態に近付くことの難しさに基づき報酬設定を行う部分である．D_{si} はエージェントが学習を開始するスタート地点の状態から i 番目の状態までの距離であり，$D_{si}avg$ は全ての状態におけるスタート地点からの距離の平均を表す．また，w_{reach} 及び w_{avg} は係数であり，それぞれ D_{si} の重み及び i 番目の状態と平均との差分の重みを表す．この項において，D_{si} の値が大きいほど大きな報酬が設定される．また，式(5.1)において後ろの 2 項である $D_{ii} \cdot w_{keep} + (D_{ii} - D_{ii}avg) \cdot w_{avg}$ は, i 番目の状態に留まることの難しさに基づき報酬設定を行う部分である．D_{ii} は i 番目の状態が自身からスタートして再び i 番目の状態に戻ってくるまでの距離であり，$D_{ii}avg$ は全ての状態についての自分自身からスタートし再びその状態へ戻るまでの距離の平均である． w_{keep} 及び w_{avg} は係数であり，それぞれ D_{ii} の重み及び i 番目の状態と平均値との差分の重みを表す.

5.2.4　実験設定

本実験におけるその他の設定を以下に示す.

■ 環境に関する設定

本実験においてエージェントが学習を行う環境は，N 種類の要素から構成される環境とし，各要素の値V_i(i=1~N)は，$\forall V_i = \{0,1\}$ の 2 値をとるものとした．また，環境の状態は各要素の値の組み合わせによって変化する．今回，環境が N 種類の要素から構成されているため，環境の状態空間は N 次元によって構成される．また，環境のとりうる状態の数 Sn は式(5.2)によって表される．

$$Sn = \prod_{i=1}^{N} 2 \tag{5.2}$$

この環境において，1 つの状態は図 5.2.2 のように表される．

環境の各状態は，Tr 種類の遷移先を持つ．今回，各状態における遷移先はランダムに設定した．このような環境でエージェントは学習を行う．

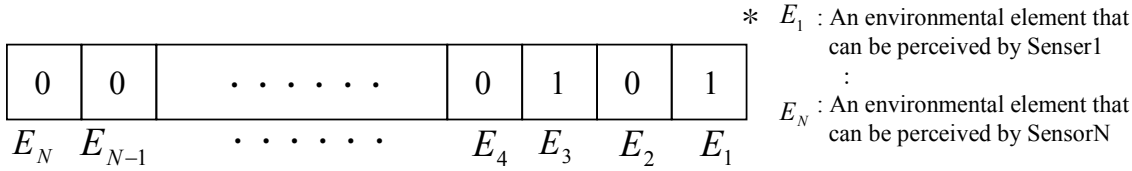


図 5.2.2　環境の状態の一例

■ エージェントに関する設定

本実験ではエージェントを以下のように設定し，学習を行わせる．エージェントの選択可能な行動は An 種類とする．またエージェントの持つセンサの種類の最大数を M 個とし，0～M 個のセンサを持つエージェントを各 1 体ずつ, 計 $M+1$ 体のエージェントに学習を行わせる．また，今回の実験で学習を行う環境は各状態に報酬が割り当てられているため，学習を行わなくてもある程度の報酬を得ることが予測される．そのため，学習が行われなかった場合の比較対象として，ランダムに行動するエージェントを 1 体作成し，ランダムに行動を行わせる．よって本実験において M +2 体のエージェントが設定した環境において行動を行う．

■ 学習手法に関する設定

本実験において，学習手法として 4 章で述べた強化学習を用いる．強化学習の各手法のうち，今回は Q 学習を用いる．また，行動の選択においては Softmax 法を用いる．

■ 学習結果の比較方法

各エージェントが学習を行った結果を比較するため，今回の実験においては学習効果を計算し，それについて比較を行うという方法を取る．l種類のセンサを持つエージェントの学習効果を$LearningEfficiency_l$とし，式(5.3)及び(5.4)により計算を行う．

$$LearningEfficiency_l = Eval_l - Eval_{random} \tag{5.3}$$

$$Eval_l = \frac{\sum_{k=1}^{T}\sum_{i=1}^{R} r_{i,S,k}}{T} \tag{5.4}$$

ここで，$Eval_l$はl種類のセンサを持つエージェントの全行動の評価を表し，$r_{i,j,k}$はk種類目の環境において，スタート地点に戻った回数j回目の，i回目の学習（行動選択）において得られた報酬を表す．また，R，S，Tはそれぞれ，スタート地点からの学習回数，1 種類の環境においてスタート地点に戻る回数，学習を行う環境の種類数を表す．

■ パラメータ設定

パラメータの設定を以下に示す．

表 5.2.1　環境の設定に関するパラメータ

N	(環境を構成する要素の数)	10
Tr	(環境の各状態における遷移先の数)	2
Sn	(環境のとりうる状態の数)	1024

表 5.2.2　エージェントに関するパラメータ

M	10
An	2

表 5.2.3　タスクに関するパラメータ

w_{reach}	2
w_{keep}	2
w_{avg}	3

表 5.2.4　学習手法に関するパラメータ

α	0.7
γ	0.6
τ	15

表 5.2.5　学習回数に関するパラメータ

R	(学習回数)	3000
S	(1種類の環境においてスタート地点へ戻る回数)	200
T	(学習を行う環境の種類)	20

5.2.5　実験結果

以上の設定で実験を行った結果を図 5.2.3 に示す.

結果より，センサの種類数が環境を構成する要素の数 N（=10）に近づくほど，学習効果が高くなっていることがわかる．また，センサの種類数が 0～4 個の場合，実際のセンサの種類数は異なるにも関わらず，学習効果についてはほとんど差が無いように見える.

0～4 種類のセンサを持つエージェントの学習効果の部分について拡大したグラフを図 5.2.4 に示す．このグラフより，0～4 種類のセンサを持つエージェントにおいては，センサの種類数に関係なく学習効果に差が出ていることがわかる.

この結果から，環境を構成する要素数よりもセンサの種類数があまりに少ない場合，学習効果がセンサの種類数と無関係に見えるようになることがわかる．これは，環境を認識するために用いることのできるセンサがあまりに少ないため，エージェントの環境認識が不完全になりすぎることが原因だと考えられる.

また，センサの種類数が 5～10 種類のエージェントの学習効果については，指数関数的に学習効果が上昇していることがわかる．これは，今回の実験において，環境の各要素の取りうる値が 2 値であるため，センサが 1 種類増えるにつれて認識できる環境の数が 2 倍になる．そのため，学習効果もそれに従って上昇したものであると考えられる.

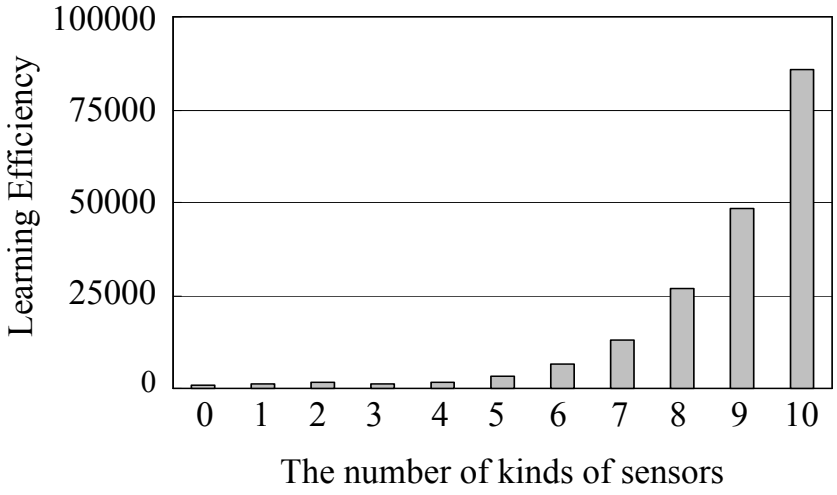


図 5.2.3　実験結果

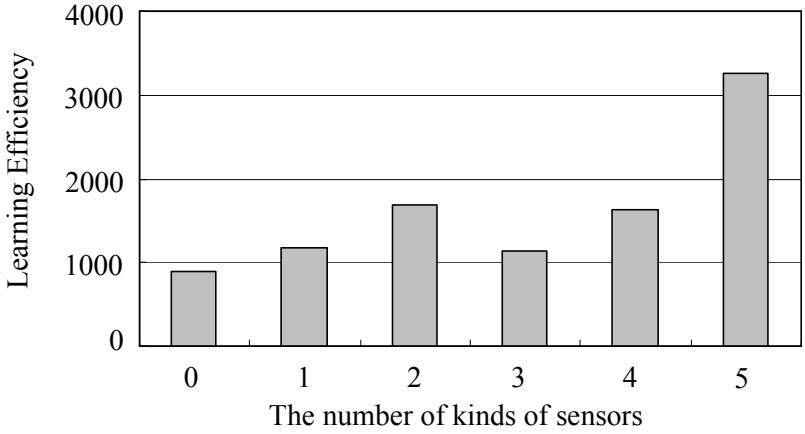


図 5.2.4　センサの種類数 0～5 のエージェントの学習結果

5.2.6　考察

今回の実験において，各エージェントは同じ学習手法及び同じ身体構造を用いて実験を行った．異なったのはセンサの種類数のみであり，センサのサンプリング周期や解像度も同じ設定である．しかし，各エージェントの学習効果については大きな違いが出る結果となった．今回の実験において，環境を構成する要素の数に比べてエージェントのセンサ種類数があまりに少ない場合，センサの種類数に関わらず学習効果が異なるという結果が出た．また，エージェントのセンサの種類数が環境を構成する要素の数に近づくにつれて，センサの種類数が増えるほど学習効果が上昇するという結果となった．

これらの結果より，ロボットに学習を行わせるという研究分野において，センサの違いによる影響を，実験を行う前に考慮することが重要であると言えるだろう．

しかし，今回の実験においては 1 種類の環境に対して，スタート地点から 3000 回移動し学習を行い，その知識を保持したままスタート地点に戻り再び学習を行うということを 200 回繰り返して行った．これは，ロボットが実環境で用いられることを考慮すると，同じ環境において 3000 回学習を行うことができ，それを 200 回繰り返すことが可能であるというのは稀なケースである．そのため，学習回数や学習結果の比較について考察を深め，実験について改善することが必要であると考えられる．

5.2.7 まとめ

今回の実験では，定常環境においてセンサの種類数の異なるエージェントに学習を行わせるという実験を行った．センサの種類数の違いにより各エージェントの学習効果に差がでることが確認された．しかし，実環境を考慮すると，実験方法に課題が残る結果となった．

5.3　定常環境における実験 (2)

本節では，エージェントが学習する環境を定常環境に設定し行った実験について述べる．また，本実験においては 5.2 で述べた課題を考慮し，エージェントの学習についてより現実の環境に近い視点で評価を行うことを目指す．

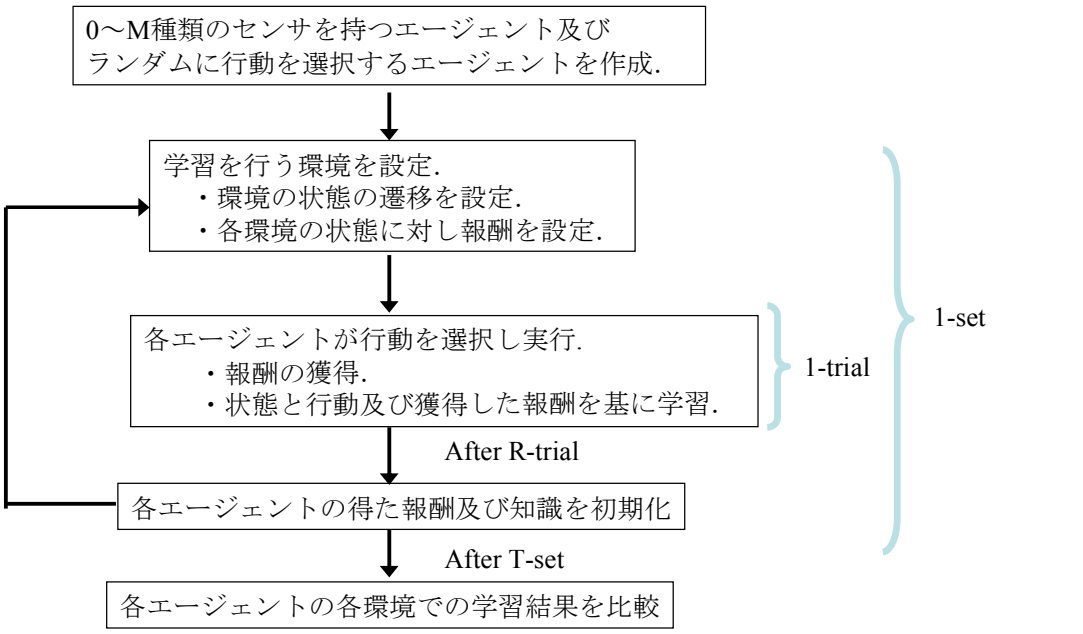
5.3.1　実験の目的

本実験では，定常環境においてセンサの種類の違いが学習効果に与える影響を調べることを目的としている．特にエージェントの学習において，学習の収束速度などの観点からロボットの学習効果についての評価を行うことを目的とする．

5.3.2　実験方法

実験方法は，4 章で述べた検証方法に基づく．N 種類の要素によって構成される環境に対して最大 M 種類（N≦M）の異なる種類のセンサを持つ複数のエージェントが学習を行うという方法を用いる．5.2 で行った実験と異なる点として，各エージェントの学習効果の評価として，最終的な報酬の総和のみならず，得られた報酬の推移から学習効果の評価を試みる．

今回の実験は次のような流れで行った．1 種類の環境において，各エージェントが R 回数学習を行う．R 回数学習を行った後，各エージェントの知識及び得た報酬を初期化し，学習を行う環境を異なるものに設定する．そして再びエージェントに学習を行わせる．今回，学習の経過を見ることに焦点を当てているため，5.2 節の実験のようにスタート地点に戻すことは行わず，代わりに R の値を大きくして実験を行うよう設定した．また，環境の種類を変えて学習を行わせるよう設定したのは，環境によってあるエージェントが学習しやすいなど傾向が出る場合が予測されるため，複数種類の環境で学習を行うことにより，環境そのものに対する全体の傾向を判定することが可能となると考えたためである．これらの流れを図 5.3.1 に示す．ここで，全てのエージェントにおける 1 回の学習を 1-trial，1 種類の環境において各エージェントが R 回の学習を完了するまでを 1-set としている．



＊ R: Trial frequency.
T: The number of kinds of environment

図 5.3.1　実験の流れ

5.3.3　実験に用いるタスク

5.2 節での実験と同様に，最大の報酬が得られるルートの探索とする．実験対象である環境の各状態へ移動することにより，エージェントは報酬を得ることができる．この報酬の総和を有限移動回数内で最大にすることがエージェントの目的となる．

ここで，環境の各状態における報酬の設定は，式(5.5)に従って設定する．

$$Rwd_i = w_{reach} \cdot D_{si} + w_{keep} \cdot \exp[\alpha_{\exp} \cdot D_{ii}] \qquad (5.5)$$

Rwd_i は i 番目の状態への移動に対し設定される報酬である．また，D_{si} はエージェントが学習を開始するスタート地点の状態から i 番目の状態までの距離であり，D_{ii} は i 番目の状態が自身からスタートしてまた i 番目の状態に戻ってくるまでの距離である．w_{reach}，w_{keep} は係数であり，それぞれ D_{si}，D_{ii} の重みを表す．$\alpha_{\exp}$ は exp の値を調整するための係数である．

この式(5.5)は，現実世界において高い報酬が得られる場所は，一般的に到達することが難しく，また留まることも難しいということを考慮し設定したものである．また，実験においてスタート地点からの移動回数が多いことから，高い報酬が得られる状態に留まることがより重要となることを考えた．そのため本実験における状態の報酬設定において，状態 i

から自分自身に戻るまでの距離について遠ければ遠いほど指数関数的に高い報酬が与えられるよう設定を行った.

5.3.4 実験設定

本実験に用いる環境, エージェント, 学習手法及び各種パラメータの設定を以下に示す.

■ 環境に関する設定

本実験においてエージェントが学習を行う環境は，5.2 の実験で設定した環境と同様である．以下に概要を示す.

本実験においてエージェントが学習を行う環境は，N 種類の要素から構成される環境とし，各要素の値V_i (i=1~N)は，$\forall V_i = \{0,1\}$ の 2 値をとるものとした．また，環境の状態は各要素の値の組み合わせによって変化する. 今回, 環境が N 種類の要素から構成されているため, 環境の状態空間は N 次元によって構成される．また，環境のとりうる状態の数を Sn とし，環境の各状態は，Tr 種類の遷移先を持つ．今回，各状態における遷移先はランダムに設定した．このような環境でエージェントは学習を行う.

■ エージェントに関する設定

本実験でのエージェントの設定は，5.2 の実験で用いた設定と同様である.

エージェントの選択可能な行動は An 種類とする．またエージェントの持つセンサの種類の最大数を M 個とし，0～M 個のセンサを持つエージェントを各 1 体ずつ，計 $M+1$ 体のエージェントに学習を行わせる．また，今回の実験で学習を行う環境は各状態に報酬が割り当てられているため，学習を行わなくてもある程度の報酬を得ることが予測される．そのため，学習が行われなかった場合の比較対象として，ランダムに行動するエージェントを 1 体作成し，ランダムに行動を行わせる．よって本実験において M +2 体のエージェントが設定した環境において行動を行う.

■ 学習手法に関する設定

本実験において，学習手法として 4 章で述べた強化学習を用いる．強化学習の各手法のうち，今回は Q 学習を用いる．また，行動の選択においては Softmax 法を用いる.

■ 学習結果の比較方法

今回の実験においては, 各エージェントの環境への適応速度を見るため, 学習を行う T 種類の環境それぞれにおいての各エージェントの報酬の総和の推移に基づき結果の比較を行う．また，1 度の移動の結果ではなく定常的に得られる報酬の量を確認するため，L 試行回

数での単純移動平均（SMA）を算出し，比較に用いる．

$$SMA = \frac{P_t + P_{t-1} + \cdots + P_{t-L-1}}{L} \tag{5.6}$$

これは，現在の時刻 T において，直近の L 個のデータの平均である．

また，学習が行われない場合（ランダムな行動選択）と学習が確実に行われる場合（センサの種類数が環境の要素と同数）と各エージェントの学習結果を比較するため，学習が行われない場合の結果を 0，確実に行われる場合の結果を 1 とし，各エージェントの総獲得報酬に対し式(5.7)のように正規化を行う．

$$RATIO_i = \frac{\sum_{k=0}^{En-1}\sum_{j=1}^{Ln} r_{i,j,k} - \sum_{k=0}^{En-1}\sum_{j=1}^{Ln} r_{random,j,k}}{\sum_{k=0}^{En=1}\sum_{j=1}^{Ln} r_{N,j,k} - \sum_{k=0}^{En-1}\sum_{j=1}^{Ln} r_{random,j,k}} \tag{5.7}$$

ここで$r_{i,j,k}$は i 種類のセンサを持つエージェントが,環境 j において,k 試行回数目に獲得した報酬の値を表す.

■ パラメータ設定

パラメータの設定を以下に示す.

表 5.3.1　環境の設定に関するパラメータ

N	(環境を構成する要素の数)	10
Tr	(環境の各状態における遷移先の数)	2
Sn	(環境のとりうる状態の数)	1024

表 5.3.2　エージェントに関するパラメータ

M	10
An	2

表 5.3.3　タスクに関するパラメータ

w_{reach}	1
w_{keep}	1
$\alpha_{\exp}$	0.65

表 5.3.4　学習手法に関するパラメータ

α	0.7
γ	0.65
τ	1000

表 5.3.5　学習回数に関するパラメータ

R　(学習回数)	120000
T　(学習を行う環境の種類)	20

表 5.3.6　学習効果の比較に関するパラメータ

L	500

5.3.5　実験結果

以上の設定で実験を行った結果を図 5.3.2～図 5.3.5 に示す.

今回実験は 20 種類の環境で学習を行ったが，結果の量が多く煩雑になるため，傾向が大きく出ている 3 環境における結果を抽出し掲載した．3 環境は便宜上環境 A，B，C と分けるが，報酬の設定方法等は同様であり，それぞれ遷移のみについて異なるものである.

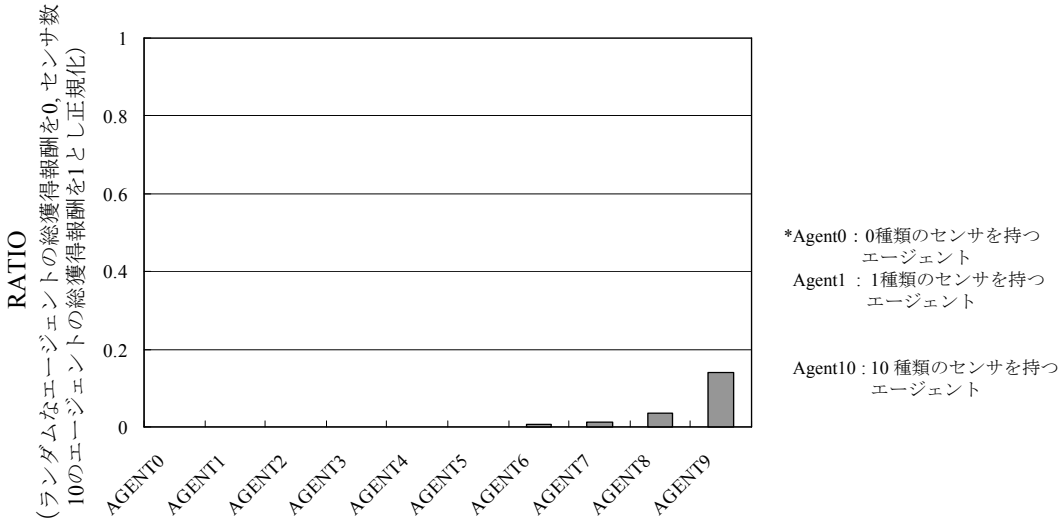
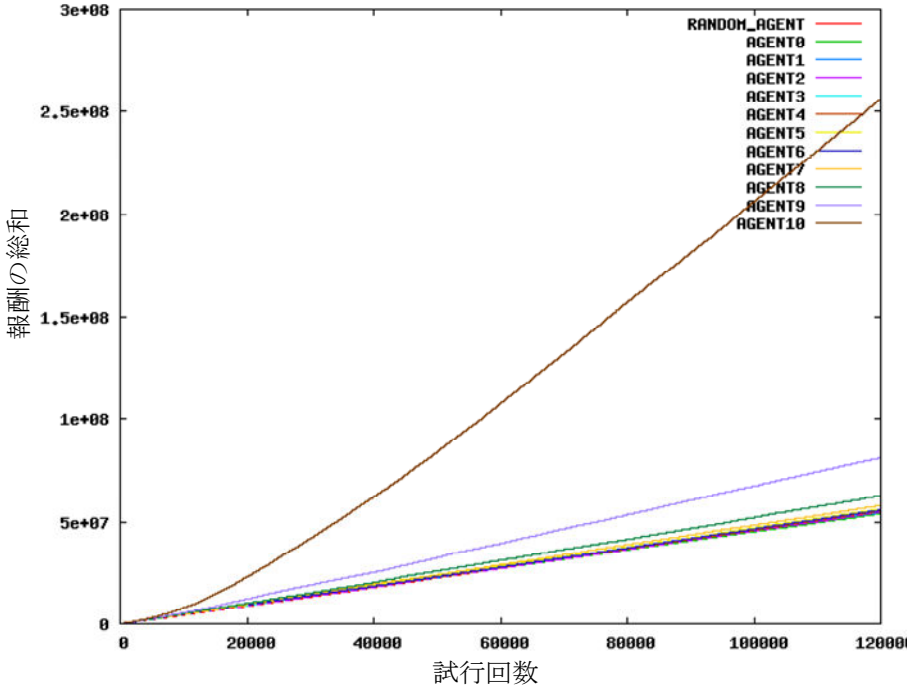
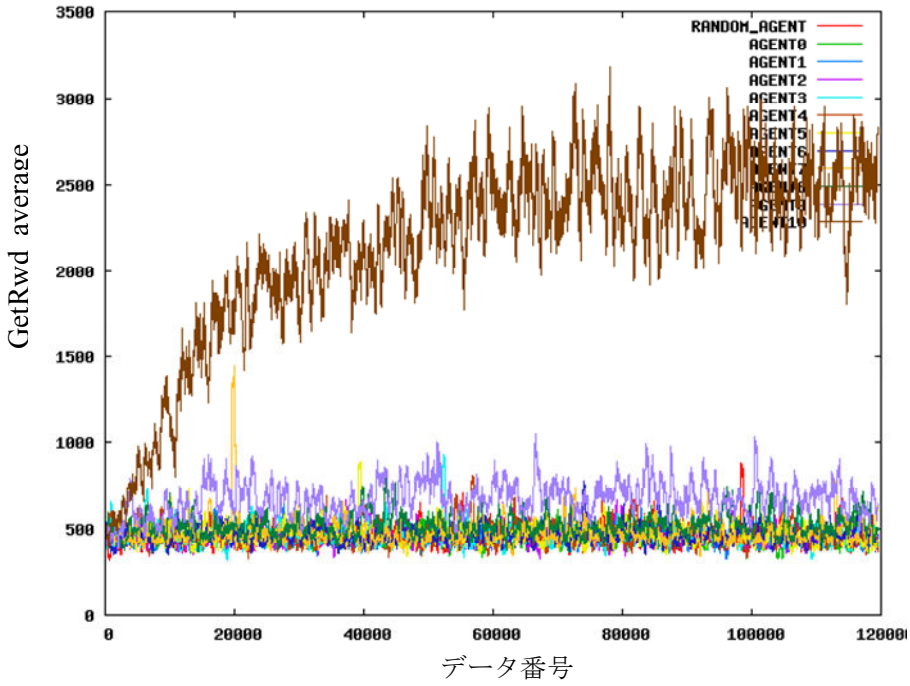


図 5.3.2　学習が行われない場合と確実に行われる場合とで正規化を行った結果

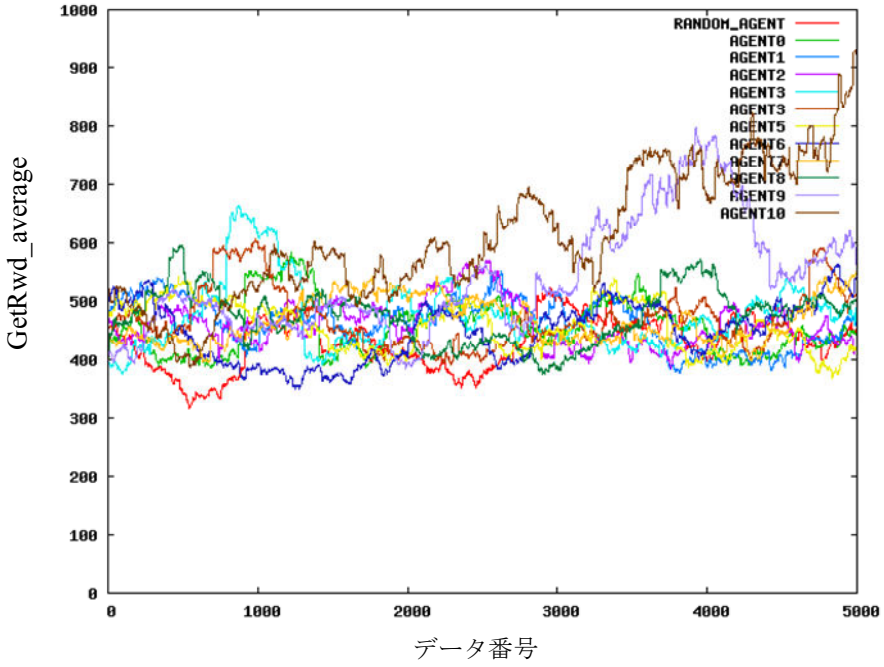
・環境 A において学習を行った結果



(a) 各エージェントの報酬の総和の推移



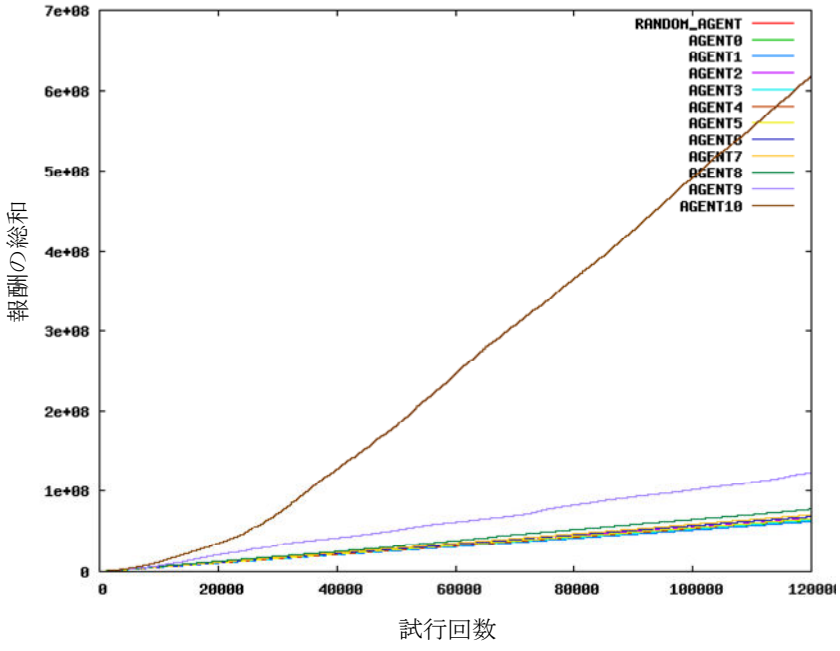
(b) L 試行回数で平均を取った報酬の推移



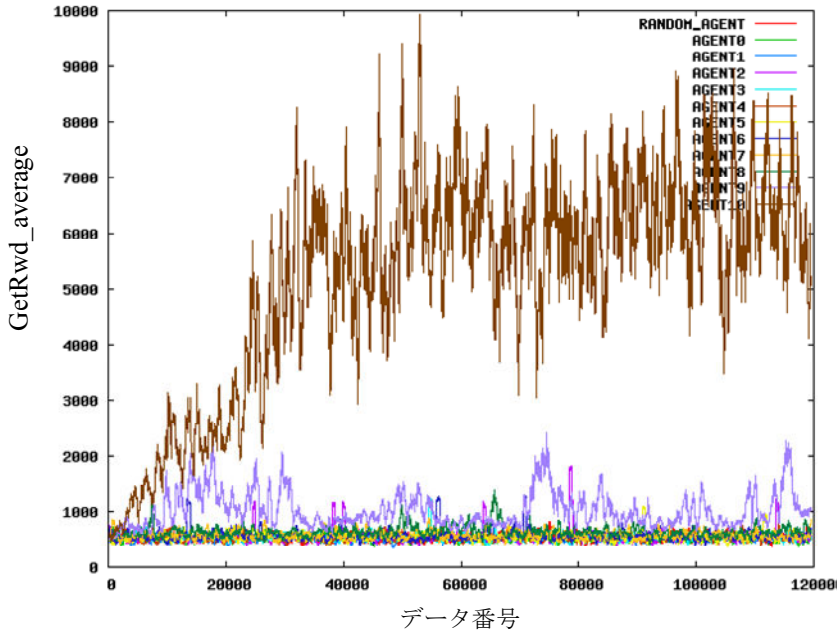
(c) 平均報酬の推移を最初の 5000 個について拡大

図 5.3.3　環境 A での各エージェントの学習結果

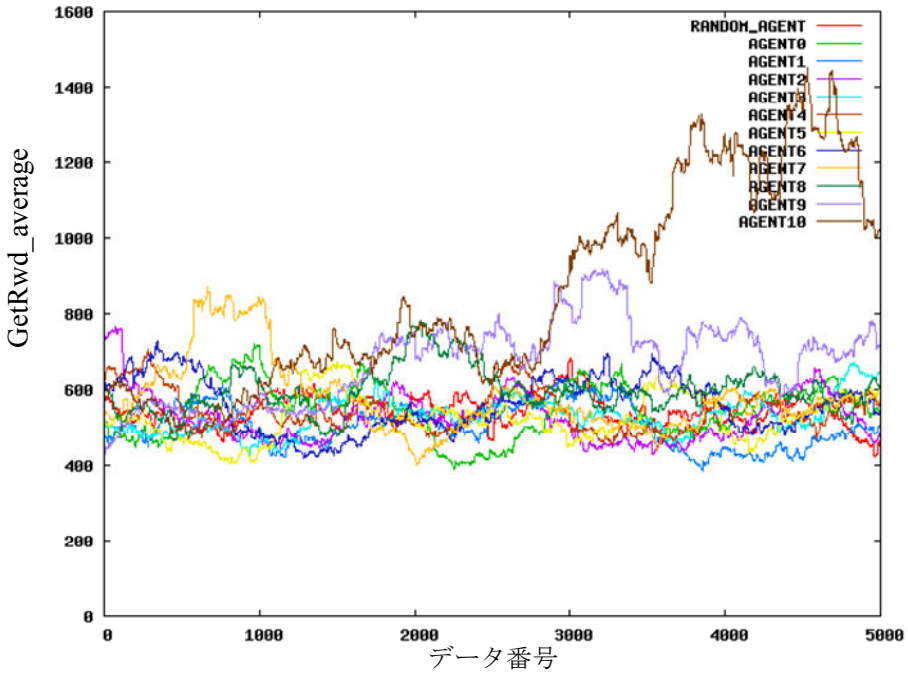
・環境 B において学習を行った結果



(a) 各エージェントの報酬の総和の推移



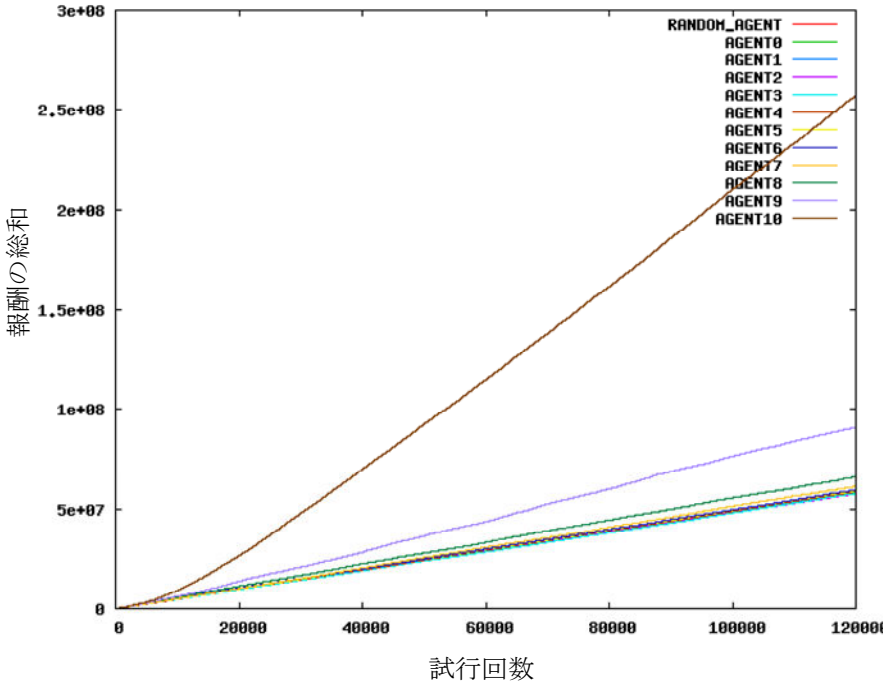
(b)　L 試行回数で平均を取った報酬の推移



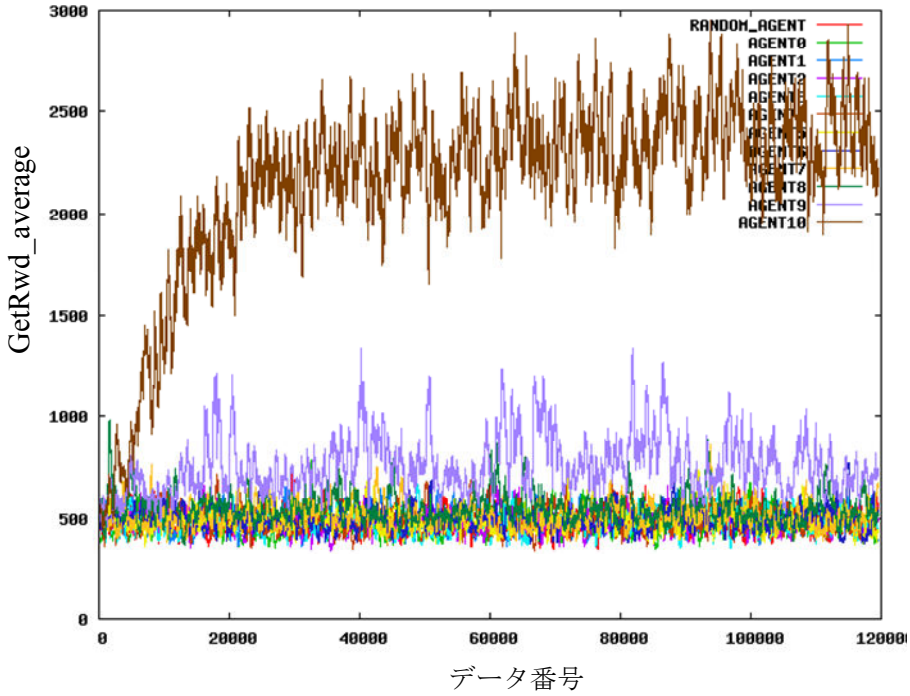
(c) 平均報酬の推移を最初の 5000 個について拡大

図 5.3.4　環境 B での各エージェントの学習結果

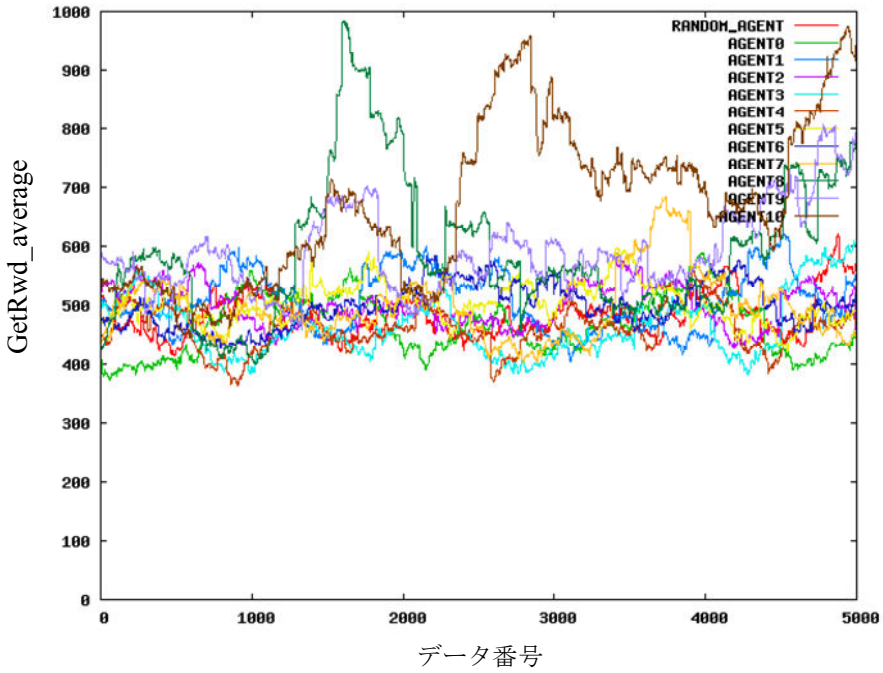
・環境 C において学習を行った結果



(a) 各エージェントの報酬の総和の推移



(b) L 試行回数で平均を取った報酬の推移



(c) 平均報酬の推移を最初の 5000 個について拡大

図 5.3.5　環境 C での各エージェントの学習結果

これらの結果より，いずれの環境においても，環境を構成する要素と同様の数，すなわち 10 種類のセンサ全てを持つエージェントの総獲得報酬が高くなるという結果が得られた．また，L 試行回数毎の平均報酬の推移から，定常的に得られる報酬も試行回数の増加と共に 10 種類のセンサを持つエージェントが最も高くなっていることがわかる．

しかし，各グラフの(c)より，平均報酬の最初の 5000 個に着目し推移を見てみると，10 種類のセンサを持つエージェントが最も高い報酬を定常的に得られるようになるまでには試行回数を重ねる必要があることがわかる．また，10 種類のセンサを持つエージェントが高い報酬を得られるようになる以前に，10 種類以下のセンサを持つエージェントの報酬が高くなっている部分があることがわかる．これは，センサの数が少ない分，環境に適した動きが早く学習できているのではないかと推測される．

5.3.6　考察

今回の実験において，5.2 節の実験と同様，各エージェントについてセンサの種類数以外は同様の設定で実験を行ったにも関わらず，各エージェントの学習効果については大きな

違いが出る結果が得られた．結果については，環境を構成する要素の数と同様の種類のセンサを持つエージェントが最も効果的に学習ができているということがわかった．しかし，学習の速度に着目して見たところ，環境の構成要素数と同種類数のセンサを持つエージェントが高い学習効果を得ることができるまでには試行回数を重ねることが必要であることが確認された．これより，ロボットを実環境に適用することを考えると，実環境は変化の多い環境であるため，学習の効果が多少低くとも速い適応性が求められる場合は，センサが多ければ良いというわけではないということが言えるだろう．

5.3.7 まとめ

今回の実験では，5.2 節の実験と異なる報酬の設定方法及び学習回数によって，定常環境でセンサの種類数の異なるエージェントに学習を行わせるという実験を行った．センサの種類数の違いにより各エージェントの学習効果に差がでることが確認され，また，学習が行われる速度についても違いがでることが確認できた．

5.4　非定常環境における実験

本節では，エージェントが学習する環境を非定常環境に設定し行った実験について述べる．

5.4.1　実験の目的

本実験では，非定常環境においてセンサの種類の違いが学習効果に与える影響を調べることを目的としている．

5.4.2　実験方法

実験方法は，3 章で述べた検証方法に基づく．N 種類の要素によって構成される環境に対して最大 M 種類の異なる種類のセンサを持つ複数のエージェントが学習を行うという方法を用いる．

今回の実験は次のような流れで行った．今回エージェントが学習を行う環境は非定常環境であるため，環境の変動を設定する．本実験において，環境の変動は次のように行った．1 種類の環境において，その環境をベースに Ch 周期で環境が変動する．これにより，各エージェントにおいて，$n \cdot Ch$ (n=自然数)回学習が行われた後に環境に変化が生じる．今回，各エージェントは同一の環境に対し学習するという方法を適用しているため，環境の変動についても，各エージェントの直面する環境はそれぞれのエージェントにおいて同様のものとなる．各エージェントは 1 種類の環境に対し R 回学習を行うものとする．R 回数学習を行った後，各エージェントの知識及び得た報酬を初期化し，学習を行う環境を異なるものに設定する．そして再びエージェントに学習を行わせる．環境の種類を変えて学習を行わせるよう設定したのは，環境によってあるエージェントが学習しやすいなど傾向が出る場合が予測されるため，複数種類の環境で学習を行うことにより，環境そのものに対する全体の傾向を判定することが可能となると考えたためである．これらの流れを図 5.4.1 に示す．ここで，全てのエージェントにおける 1 回の学習を 1-trial，1 種類の環境において各エージェントが R 回の学習を完了するまでを 1-set としている．

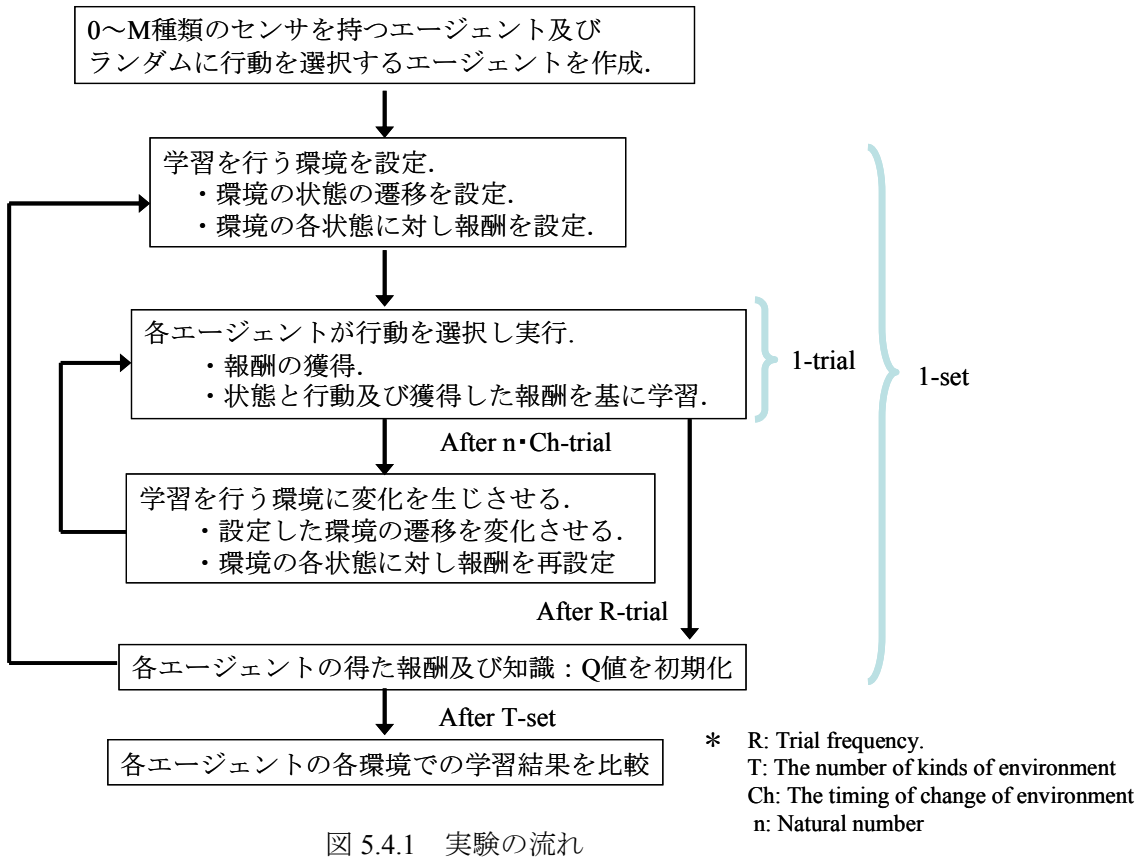


図 5.4.1　実験の流れ

5.4.3　実験に用いるタスク

本実験においてエージェントの行うタスクは，最大の報酬が得られるルートの探索である．実験対象である環境の各状態へ移動することにより，エージェントは報酬を得ることができる．この報酬の総和を有限移動回数内で最大にすることがエージェントの目的となる．

ここで，環境の各状態における報酬の設定は，5.3 の定常環境での実験(2)と同様に式(5.5)に従って設定する．

5.4.4　実験設定

本実験における環境，エージェント及びパラメータの設定を以下に示す．

■ 環境に関する設定

本実験においてエージェントが学習を行う環境は，N 種類の要素から構成される環境とし，各要素の値 V_i (i=1~N)は，$\forall V_i = \{0,1\}$ の 2 値をとるものとした．また，環境の状態は各要素の値の組み合わせによって変化する．

本実験において，環境の取りうる状態の数は，各要素の値の取りうる組み合わせの数であり Sn とする．環境の各状態は，Tr 種類の遷移先を持つ．今回，各状態における遷移先はランダムに設定した．このような環境でエージェントは学習を開始する．

また本実験は非定常環境における実験であるので，環境に変動を起こす必要がある．環境の変動は今回次のようになる．各エージェントが Ch 回試行を行った後，環境の各状態から次の状態へのリンクに対し，P %の割合でリンク先を変化させるよう設定した．

リンク先の変更に伴い，各状態についてスタート地点からの距離及び状態 i からスタートして再び状態 i に戻るまでの距離に変化が生じる．そこで，各状態に設定される報酬を再び設定し直す必要がある．報酬の再設定は，式(5.8)に従って設定を行う．

$$Rwd_i = w_{keep} \cdot \exp[\alpha_{\exp} \cdot D_{ii}] \qquad (5.8)$$

式(5.8)において，Rwd_i は i 番目の状態への移動に対し設定される報酬である．また，D_{ii} は i 番目の状態が自身からスタートしてまた i 番目の状態に戻ってくるまでの距離である．w_{keep} は係数であり D_{ii} の重みを表す．$\alpha_{\exp}$ は exp の値を調整するための係数である．

環境を変動させる際，各エージェントはスタート地点から他の状態へ移動していることが考えられる．そのため，報酬の再設定においてスタート地点からの距離を考慮に入れず，状態 i からスタートしそこへ再び戻るまでの距離のみを考慮に入れた．

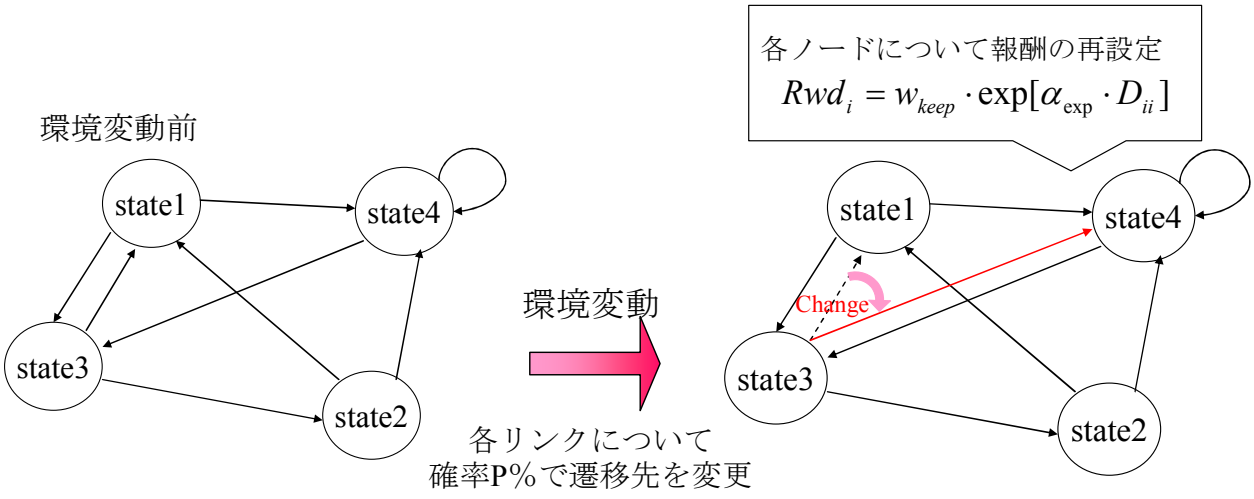


図 5.4.2　環境変動の例：状態数 4 の場合

■ エージェントに関する設定

本実験ではエージェントを以下のように設定し，学習を行わせる．エージェントの選択可能な行動は An 種類とする．またエージェントの持つセンサの種類の最大数を M 個とし，0～M 個のセンサを持つエージェントを各 1 体ずつ, 計 $M+1$ 体のエージェントに学習を行わせる．また，今回の実験で学習を行う環境は各状態に報酬が割り当てられているため，学習を行わなくてもある程度の報酬を得ることが予測される．そのため，学習が行われなかった場合の比較対象として，ランダムに行動するエージェントを 1 体作成し，ランダムに行動を行わせる．よって本実験において M +2 体のエージェントが設定した環境において行動を行う．

■ 学習手法に関する設定

本実験において，学習手法として 4 章で述べた強化学習を用いる．強化学習の各手法のうち，今回は Q 学習を用いる．また，行動の選択においては Softmax 法を用いる．

■ 学習結果の比較方法

今回の実験においては，各エージェントの環境への適応速度を見るため，学習を行う T 種類の環境それぞれにおいての各エージェントの報酬の総和の推移に基づき結果の比較を行う．また，5.3 節での定常環境での実験(2)と同様，式(5.5)により，L 試行回数の平均を算出し，比較に用いる．また，式(5.7)により環境を確実に認識できる場合と学習が行われない場合とで正規化を行い，結果の検討に用いる．

■ パラメータ設定

パラメータの設定を以下に示す．

表 5.4.1　環境の設定に関するパラメータ

N	(環境を構成する要素の数)	10
Tr	(環境の各状態における遷移先の数)	2
Sn	(環境のとりうる状態の数)	1024

表 5.4.2　環境の変化に関するパラメータ

Ch	2500
P	10

表 5.4.3　エージェントに関するパラメータ

M	10
An	2

表 5.4.4　タスクに関するパラメータ

w_{reach}	1
w_{keep}	1
$\alpha_{\exp}$	0.65

表 5.4.5　学習手法に関するパラメータ

α	0.7
γ	0.65
τ	1000

表 5.4.6　学習回数に関するパラメータ

R　(学習回数)	50000
T　(学習を行う環境の種類)	20

表 5.4.7　学習効果の比較に関するパラメータ

L	500

5.4.5　実験結果

以上の設定で実験を行った結果を図 5.4.3～5.4.6 に示す.

今回実験は 20 種類の環境で学習を行ったが，結果の量が多く煩雑になるため，傾向が大きく出ている 3 環境における結果を抽出し掲載した．3 環境は便宜上環境 A，B，C と分けるが，報酬の設定方法等は同様であり，それぞれ遷移のみがランダムに選択されるため異なるものである.

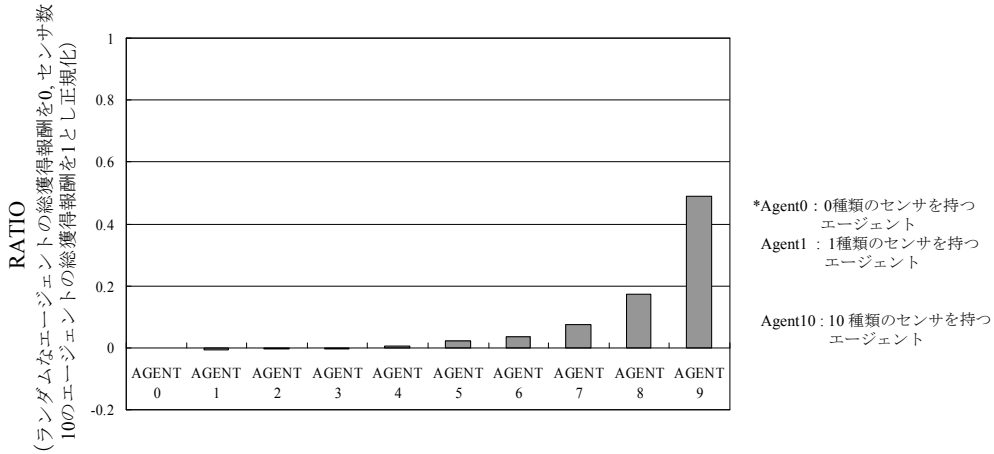
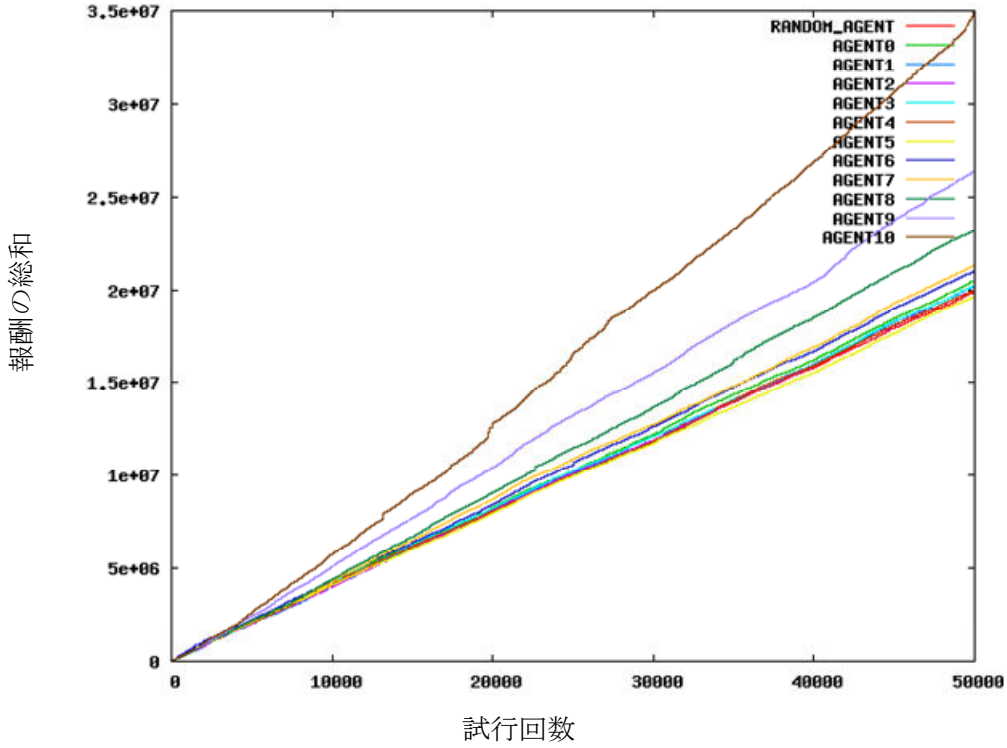
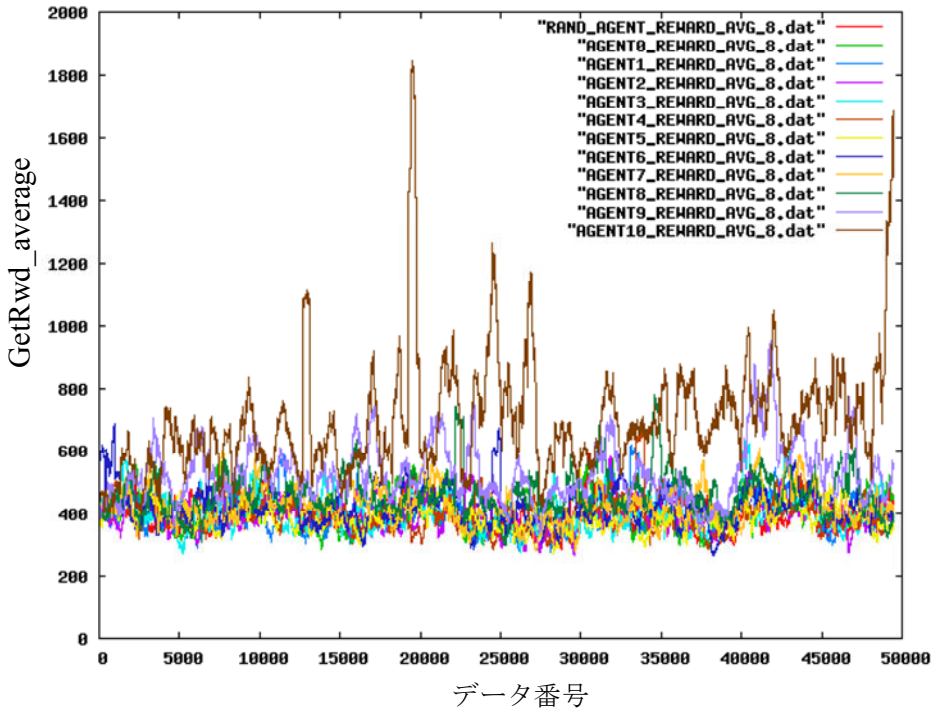


図 5.4.3　学習が行われない場合と確実に行われる場合とで正規化を行った結果

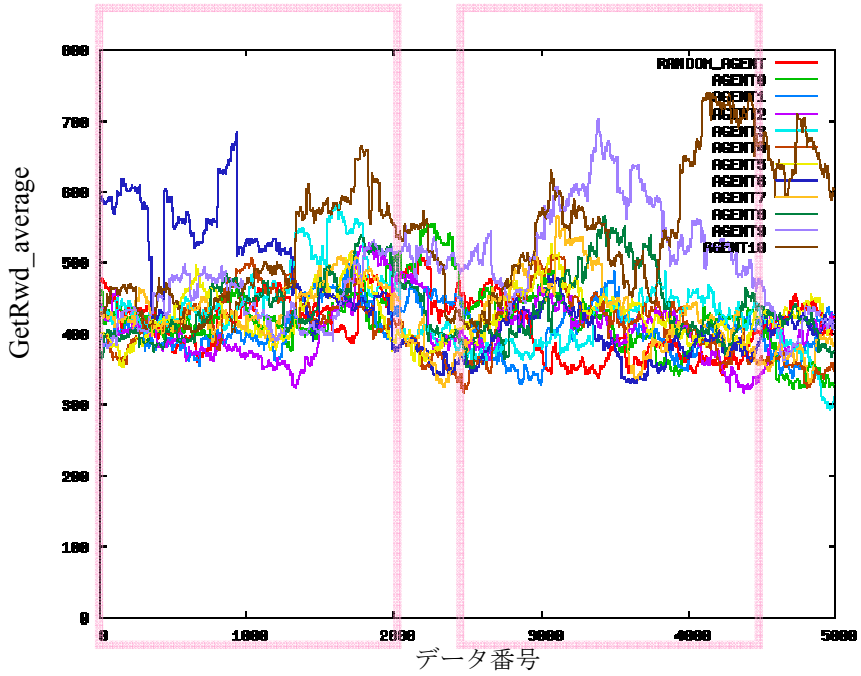
・環境 A において学習を行った結果



(a) 報酬の総和の推移



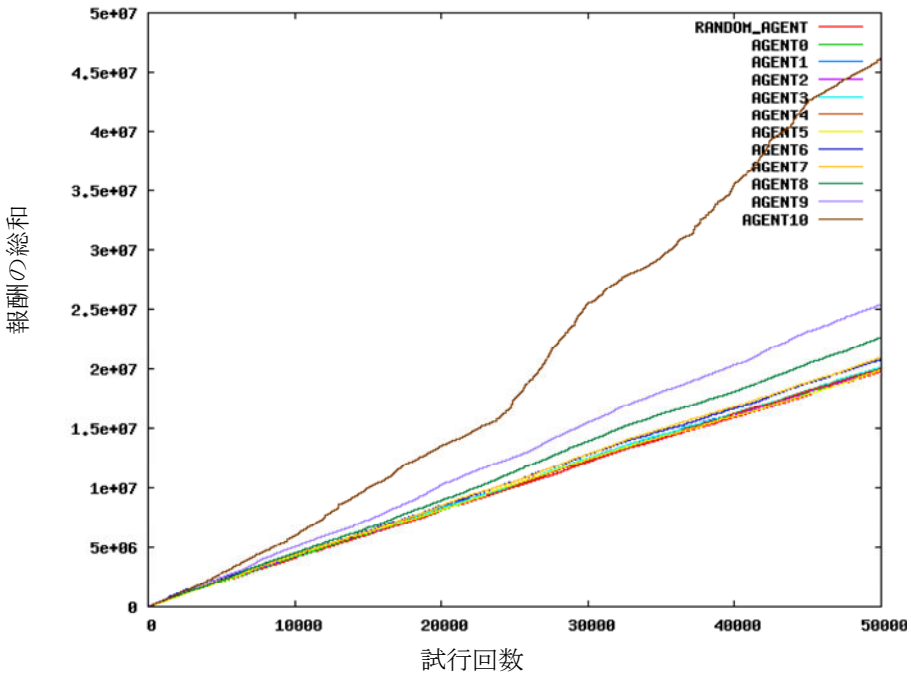
(b) L 試行回数で平均を取った報酬の推移



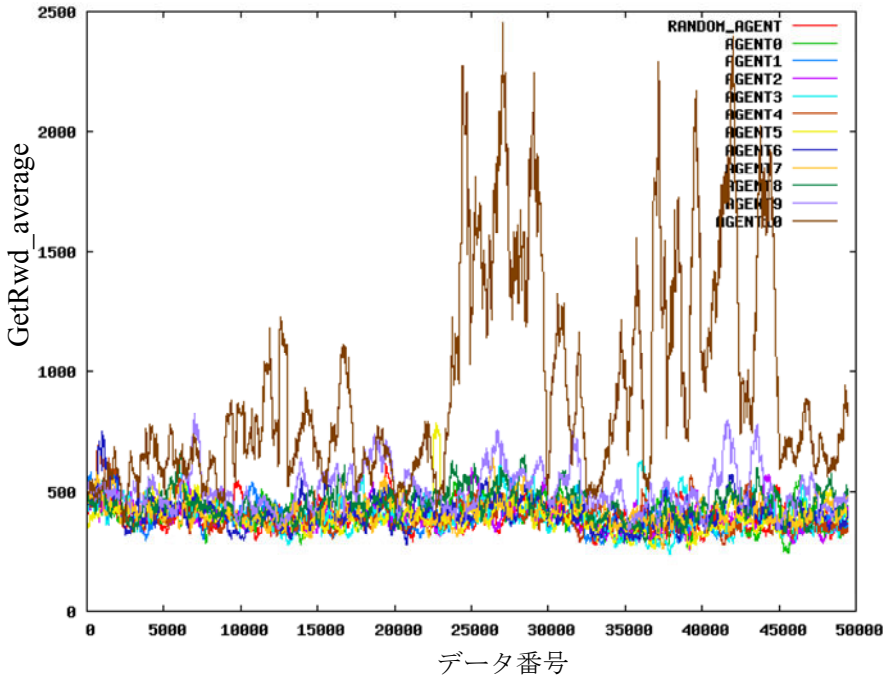
(c) L 試行回数で平均を取った報酬の推移を最初の 5000 個について拡大

図 5.4.4　環境 A での各エージェントの学習結果

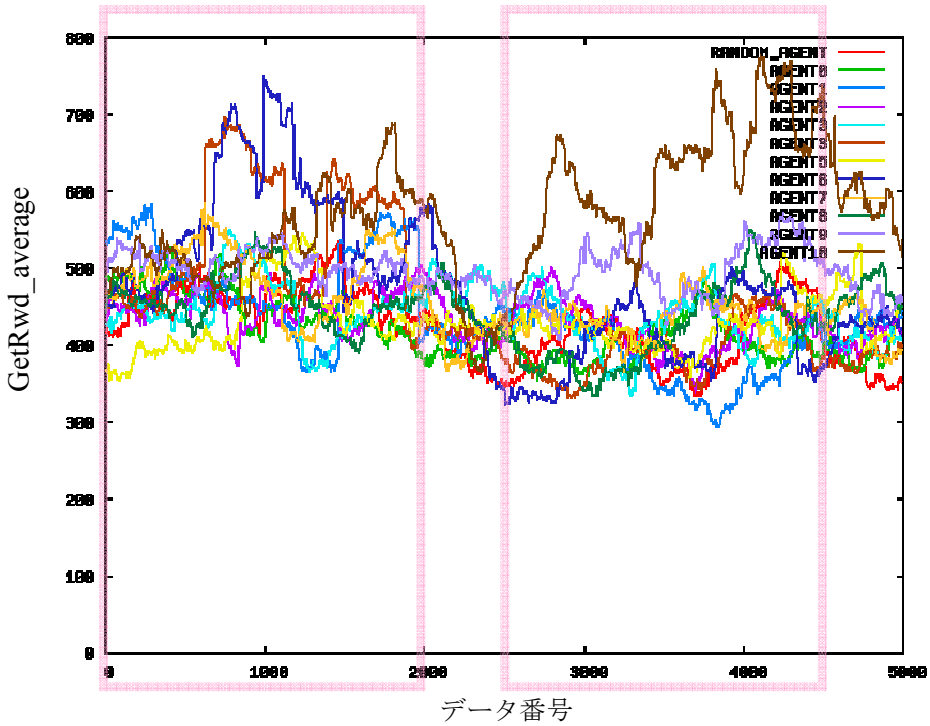
・環境 B において学習を行った結果



(a) 報酬の総和の推移



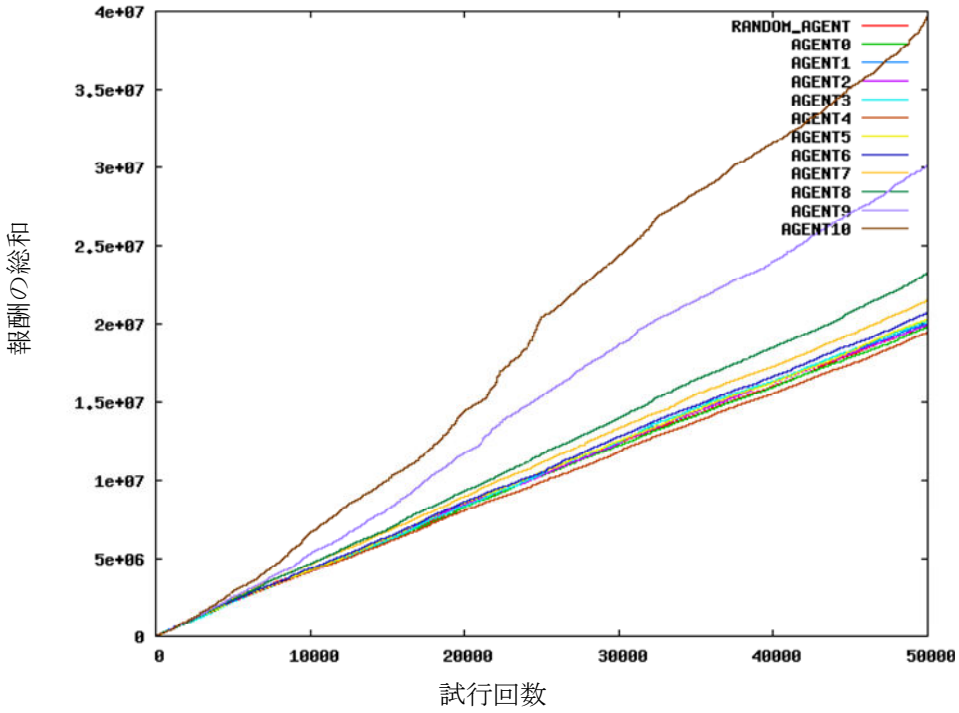
(b) L 試行回数で平均を取った報酬の推移



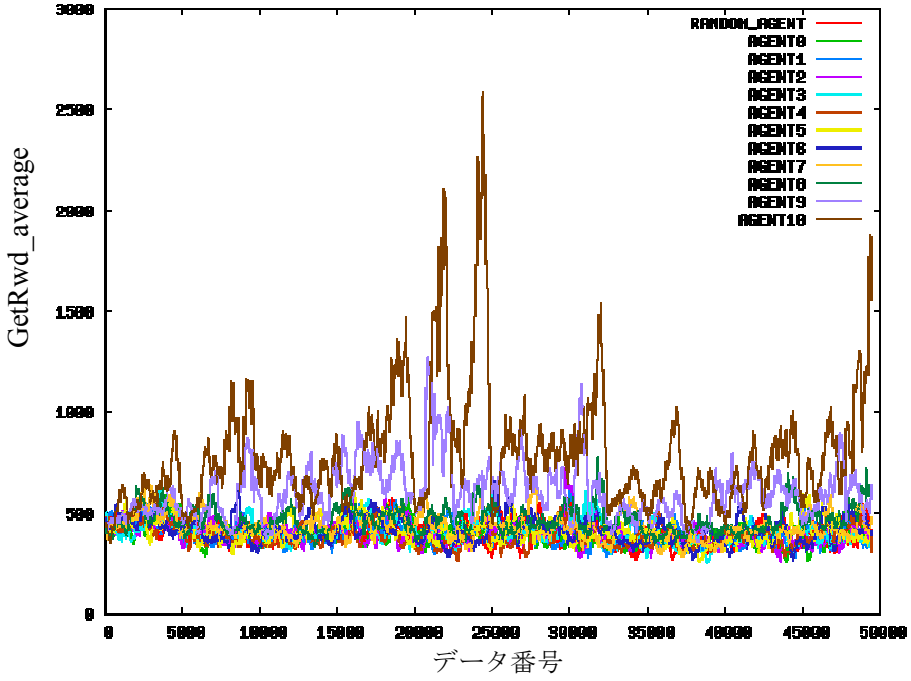
(c) L 試行回数で平均を取った報酬の推移を最初の 5000 個について拡大

図 5.4.5　環境 B での各エージェントの学習結果

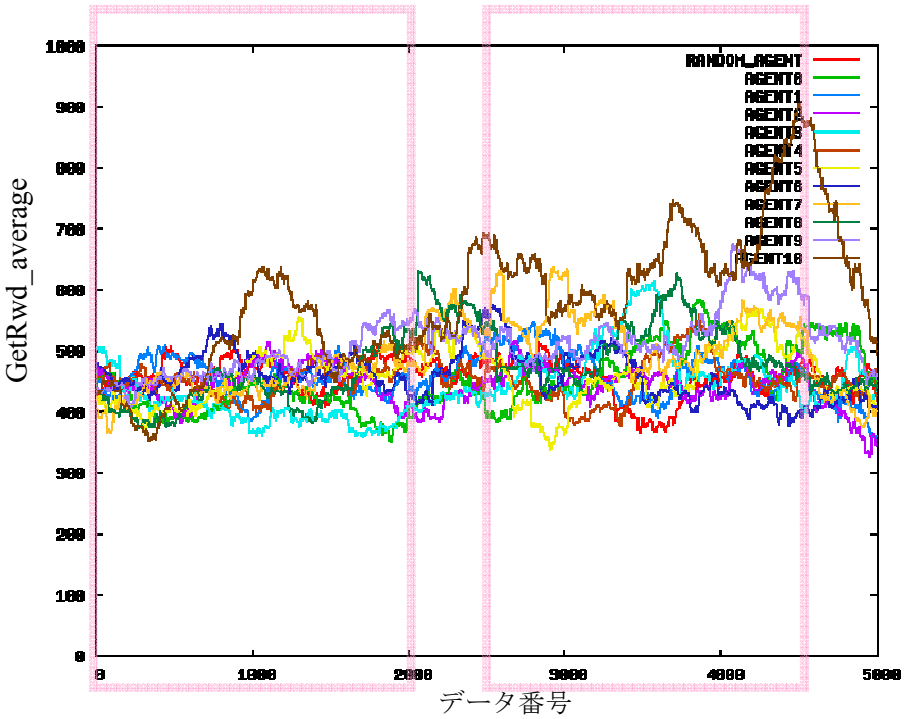
・環境 C において学習を行った結果



(a) 報酬の総和の推移



(b) L 試行回数で平均を取った報酬の推移



(c) L 試行回数で平均を取った報酬の推移を最初の 5000 個について拡大

図 5.4.6　環境 C での各エージェントの学習結果

これらの結果より，いずれの環境においても，環境を構成する要素と同様の数，すなわち 10 種類のセンサ全てを持つエージェントの総獲得報酬が高くなるという結果が得られた．また，L 試行回数毎の平均報酬の推移から，定常的に得られる報酬も 10 種類のセンサを持つエージェントのものが総合的に見て高いという傾向が見られた．しかし，定常的に得られる報酬においては，常に 10 種類のセンサ全てを持つエージェントのものが高いというわけではない．これは，環境が 2500 回毎に変動するため，変動した環境に適応するのに時間がかかっているためであると考えられる．

各グラフの(c)より，平均報酬の最初の 5000 個に着目し推移を見てみる．今回 500 回移動平均をとっているため，グラフにおいて 0 の部分が最初の 500 試行の平均値となる．また，環境の変動後 500 回は前の環境において得た報酬が値に影響すると考えられる．このことより，各グラフにおいてピンク色の枠で囲っている部分が環境 A~C において変動前に定常的に獲得した報酬及び 1 回の変動後に定常的に得られた報酬となる．それぞれピンク色で囲った部分に着目すると，学習初期のうちはセンサの種類数が 10 であるエージェントの獲得報酬は必ずしも高くなく，センサの種類数が少ないエージェントの獲得報酬が高い場合も多々あることが確認できる．そして試行回数の増加に伴い，10 種類のセンサを持つエージェントが定常的に得る報酬が高くなっている傾向が確認できる．

5.4.6 考察

今回の実験において，各エージェントについてセンサの種類数以外の要素，センサのサンプリング周波数や分解能及び用いる学習手法やロボットの身体構造については同様の設定で実験を行ったにも関わらず，各エージェントの学習効果については差が出る結果となった．結果については，非定常環境においても環境を構成する要素の数と同様の種類のセンサを持つエージェントが最も効果的に学習ができているということがわかった．これは今回の実験において非定常環境の設定は完璧にランダムに遷移するものではなかったため，ある程度の傾向を学習することが可能であったためであると考えられる．しかし環境が変動した場合，環境の構成要素数と同種類数のセンサを持つエージェントが高い報酬を得られるようになるにはある程度の試行回数が必要となることが確認された．また，環境の構成要素数と同種類数のセンサを持つエージェントが十分な試行を行う前にセンサの少ないエージェントが高い報酬を得る場合も確認できた．これより，ロボットを実環境に適用することを考えると，実環境は変化の多い環境であるため，学習の効果が多少低くとも速い適応性が求められる場合は，センサが多ければ良いというわけではないということが言えるだろう．

5.4.7 まとめ

今回の実験では，非定常環境においてセンサの種類数の異なるエージェントに学習を行わせるという実験を行った．センサの種類数の違いにより各エージェントの学習効果に差が出ることが確認された．また，各エージェントが非定常環境に適用する速度に違いが生じることも確認できた．

5.5 考察

本節では，本章で行った実験全体の考察を行う．

5.2 節での実験では，定常環境において，同じ学習手法及び同じ身体構造を持つ複数のエージェントにおいて，センサの種類数のみを変えて実験を行った．その結果，各エージェントの学習効果について大きな違いが出るという結果が得られた．また，エージェントのセンサの種類数が環境を構成する要素の数に近付くにつれてセンサの種類数と学習効果が比例関係にあるということが確認された．また，環境を構成する要素の数に比べてエージェントのセンサの種類数があまりに少ない場合，センサの種類数と学習効果の間に関係性は見受けられず，センサの種類数が M のエージェントの学習効果がセンサの種類数 M+δ (δ>0)のエージェントの学習効果よりも高い場合もあるという結果が確認された．これらのことより，環境の構成する要素の数に対しセンサの種類数があまりに少ない場合は，認識できる環境の要素数が多少増えたところで学習効果に差は出ないということが考えられる．

5.3 節では，定常環境において，5.2 節で用いたエージェントと同じ複数のエージェントを用いて学習を行わせた．ただし学習の過程を見るため，実験の回数を変更して学習を行わせた．その結果，環境を構成する要素の数（＝N）と同様の種類数のセンサを持つエージェントの総獲得報酬が最も高いという結果が得られた．また，定常的に得られる報酬も試行回数の増加と共に N 種類のセンサを持つエージェントの値が最も高くなっていることが確認された．しかし，学習の過程に着目すると，N 種類のセンサを持つエージェントが最も高い報酬を定常的に得られるようになるまでには試行回数を重ねる必要があるということが確認された．また，学習の初期段階を見ると，N 種類のセンサを持つエージェントが高い報酬を得られるようになる以前に，N 種類以下のセンサを持つエージェントが定常的に得られる報酬が高くなっていることがわかった．これより，センサの数が少ない分認識する状態の数が少なく，そのため環境の傾向が掴めた場合，環境に適した動きが早く学習できるのではないかということが推測された．また，ロボットを実環境に適用することを考えた場合，実環境は変化の多い環境であるため，学習の効果が多少低くとも速い適応性が求められる場合も多いと考えられる．その場合，センサの種類数が多ければ必ずしも良いとは言えないということが考察された．

5.4 節では，非定常環境において，5.2 節及び 5.3 節で用いたエージェントと同じ設定の複数のエージェントを用いて学習を行わせた．その結果，非定常環境においても，環境を構成する要素の数（＝N）と同様の種類数のセンサを持つエージェントの総獲得報酬が最も高いという結果が得られた．しかし，定常的に得られる報酬の推移を見ると，N 種類のセンサを持つエージェントが定常的に得られる報酬が総合的に最も高い結果であるが，定常環境における実験の場合と異なり，その振れ幅が大きくなっている．これは，環境の変動に追従するのに時間がかかったがためであると考えられた．また，環境が変動した場合，N 種類

のセンサを持つエージェントが定常的に高い報酬をえら得るようになるにはやはり試行回数を重ねることが必要となることが確認された. N 種類のセンサを持つエージェントが十分な試行回数を重ねる前に N 種類以下のセンサを持つエージェントが定常的に得られる報酬が高くなる場合も確認された．これより，実際に非定常環境で実験を行った結果，やはり高い環境追従性が必要とされる場合は，センサの種類が必ずしも環境の要素数に対応していれば良いわけではないことが考えられる．

これより，以下のことが今回の実験によって考えられる．

- 試行回数が無限にとれる学習であれば，環境の構成要素数 N と同等のセンサの種類を持つエージェントの学習効果が最も高い．
- エージェントの持つセンサの種類数が N に対してあまりに少ない場合，認識できる環境の要素数が多少増えたところで学習効果に差は出ず，低い部分に留まる場合が考えられる．
- 環境追従性を考えると，エージェントのセンサの種類数が N より少ない場合の方が素早く環境に適応できる可能性がある．
- 有限試行回数の学習においては，必ずしもセンサの種類数が N である場合が良いとは限らない．

これらのことより，ロボットに学習を行わせるという研究分野において，センサの種類数の違いによる影響を，実験を行う前に考慮することが重要となることが指摘されると考えられる．

5.6 まとめ

本章では，センサの種類数の違いが学習効果に及ぼす影響についての検証実験を行った．ロボットが学習を行う環境を定常環境と非定常環境の 2 種類に設定し，それぞれ実験を行った．また，定常環境での実験において，報酬の与え方について異なる 2 種類の設定方法を用いそれぞれ実験を行った．実験の結果を示し，それに基づきロボットのセンサの種類数の違いによる環境認識能力の差が学習効果に及ぼす影響について確認した．

第 6 章　結論

6.1　まとめ

ロボットの環境認識能力の違いが学習効果に及ぼす影響について調べることを目標とした．ロボットの環境認識能力に影響を及ぼすセンサの要素として，センサの種類，分解能，サンプリング周波数を考え，それぞれがロボットの環境認識能力にどのように関係し，学習効果に与えうる影響について考察を行った．本論文においては特にセンサの種類数に着目し，センサの種類数の違いが学習効果に及ぼす影響について調べることを目的とした．

学習効果への影響を調べる方法として，センサの種類数の異なるエージェントが同一の環境において学習を行い，その結果を比較するという方法を用いた．この方法に基づきシミュレーションを用いて実験を行った．実験においては，ロボットが学習を行う環境を変動のない定常環境と，変動のある非定常環境に分け各環境で実験を行った．

実験の結果として，センサの種類数が環境を構成する要素の数と同じ場合が最も高い学習効果が得られているということが確認された．また，環境への適応速度を見ると，センサの種類数が多いエージェントは高い報酬が得られるまでに多くの試行回数を必要としていることが確認された．これに対しセンサの種類数の少ないエージェントは，最も高くはないがある程度高い報酬が得られるようになるまでに，センサの種類数の多いエージェントよりも短い試行回数で至ることができる傾向が見られた．これより，ロボットが変化の多い動的な環境で用いられることを考えると，必ずしもセンサの種類数が多い場合が良いとは言えないことが予測された．

6.2　今後の課題

・実験について

今回の実験において，非定常環境の変動の設定は，全てのエージェントが Ch 試行回数学習を行った後，各ノードの各リンクについて P%の確率で遷移先を変更するというものであった．しかし，非定常環境には他にエージェントの行動が環境自体に変化を及ぼす場合や，時系列的に環境が変化する場合などが考えられる．そのため，それらについても実験を行うことが今後必要であると考える．また，今回の実験は全てシミュレーションで行った．しかしロボットが実体を持ち実環境において用いられることを考えると，実機における実

験も必要となることが考えられる．

・対象とするロボットの環境認識能力について

本論文においては，特にセンサの種類について着目し実験，考察等を行った．しかし，前述のとおり，ロボットの環境認識能力に影響を与えるセンサの要素としては，他にセンサの分解能やサンプリング周波数等が考えられる．そのため，それらについても実験を行うことが今後の課題の 1 つとして挙げられる．

・学習効果に影響を及ぼす要素について

本論文においては特にロボットの環境認識能力に着目し，ロボットのセンサを対象として学習効果に及ぼす影響を考察した．しかし，ロボットの学習においては，他に用いる学習手法やロボットの身体構造なども学習効果に影響を与えるものとして考えられる．そのため，今後それらについても同様に考察を行い，学習効果に及ぼす影響を調べる必要性があると考えられる．

謝辞

本論文を結ぶにあたり，日頃から様々な面で有益な御指導・御助言をいただきました主指導教員の倉重健太郎先生に深く感謝の意を表します．また，中間発表や学会発表練習の場で御指導，御助言を下さいました本学情報工学科の畑中雅彦先生，本田泰先生，佐賀聡人先生，鈴木幸司先生，前田純治先生，魚住超先生，須藤秀紹先生，渡部修先生，渡邉真也先生，インタラクティブシステム研究室の西川玲さんにこの場をお借りして厚くお礼を申し上げます．また，研究活動全般において有意義なディスカッションや意見をいただいた認知ロボティクス研究室の池田憲弘君，木島康隆君，池田善治君，黒滝麗子さんに心より感謝致します.

参考文献

[1]Masato Hirose, Kenichi Ogawa, “Honda humanoid robots development”, Philosophical Transactions of The Royal Society A, Vol.364, No.1850, pp.11-19, 2007.

[2]柴田崇徳,”人とロボットの身体的インタラクションを通した主観的価値の創造―アザラシ型メンタルコミットロボットの研究開発―”,日本ロボット学会,Vol.18, No.2, pp.200-203, 2000.

[3]Tobias Ramforth, “History of Robotics Development”, Seminar Paper(Seminar Human-Robot Interaction Universitat Dortmund), pp.8-19, 2006.

[4]藤田善弘, “チャイルドケアロボット PaPeRo”, 日本ロボット学会誌, Vol.24, No.2, pp.162-163, 2006.

[5] Jun Tani, “Model-based learning for mobile robot navigation from the dynamic system perspective”, IEEE Trans. on Systems, Man, and Cybernetics Part B: Cybernetics, Vol.26, No.3, pp.421-436, 1996.

[6] 木村元, 宮崎和光, 小林重信, “強化学習システムの設計指針”, 計測と制御, Vol.38, No.10, pp.618-623, 1999.

[7] Yasutake Takahashi, Minoru Asada, “Multilayered learning systems for vision-based behavior acquisition of real mobile robot”, Proceedings of SICE Annual Conference 2003 in Fukui, pp.2937-2942, 2003.

[8] Naoyuki Kubota, Daisuke Hisajima, Fumio Kojima, Toshio Fukuda, “Fuzzy and Neural Computing for Communication of a Partner Robot”, Journal of Multiple-valued Logic and Soft-Computing, Vol.9, No.2, pp.221-239, 2003.

[9] 前田隆, 青木文夫, “新しい人工知能：基本編”, オーム社, 1999 年.

[10] Stuart Russell, Peter Norving, “Artificial Intelligence: A Modern Approach (Second Edition)”, Prentice Hall, 2002.

[11] 銅谷賢治, “計算神経科学への招待: 脳の学習機構の理解を目指して”, サイエンス社, 2007.

[12] 安居院猛, 長橋宏, 高橋裕樹, “ニューラルプログラム”, 昭晃堂, 1993.

[14] 岡田慧, “日常生活支援ヒューマノイドの環境認識・行動制御”, 日本ロボット学会誌, Vol.26, No.4, pp.330-333, 2008.

[15] 鈴木太郎, 目黒淳一, “小型自律飛行ロボットを用いた災害時における情報収集システムの構築”, 日本ロボット学会誌, Vol.26, No.6, pp.553-560, 2008.

[16] Emil M. Petriu, Thom E. Whalen, Rami Abielmona, Alan Stewart, “Robotic Sensor Agents : A new generation of intelligent agents for complex environment monitoring.”, Instrumentation & Measurement Magazine, IEEE, vol.7, Issue 3, pp.46-51, 2004.

[17] “特集：ロボットのデザイン”, 季刊 d/sign, No.13, pp.14-36, 2006.

[18] “解説：プロトタイプロボット展のために開発された全てのロボットの紹介”, 日本ロボット学会誌, Vol.24, No.2, pp.171-204, 2006.

[19] Kentarou Kurashige, Yukiko Onoue, “The robot learning by using sense of pain”, Proceeding of International Symposium of Humanized Systems 2007, pp.1-4, 2007.

[20] 矢野雅文, 富田望, “実環境における 2 足歩行の創発的リアルタイム制御”, 日本ロボット学会誌, Vol.23, No.1, pp.11-16, 2005.

[21] 港隆史, 浅田稔, “環境の変化に適応する移動ロボットの行動獲得”, 日本ロボット学会誌, Vol18, No.5, pp.706-712, 2000.

[22] Richard. S. Sutton, Andrew. G. Barto, “Reinforcement Learning”, The MIT Press, 1998.

[23] 高橋泰岳, 浅田稔, “実ロボットによる行動学習のための状態空間の漸近的構成”, 日本ロボット学会, Vol.17, No.1, pp.118-124, 1999.

[24] 木村元, 山下透, 小林重信, “強化学習による 4 足ロボットの歩行動作獲得”, 電気学会電子情報システム部門誌, Vol.122-C, No.3, pp.330-337, 2002.

[25] 浅田稔, 野田彰一, 俵積田健, 細田耕, “視覚に基づく強化学習によるロボットの行動獲得”, 日本ロボット学会誌, Vol.13, No.1, pp.68-74, 1995.

[26] 浅田稔, “強化学習の実ロボットへの応用とその課題”, 人工知能学会誌, Vol.12, No.6, pp.831-836, 1997.

[27] Yukiko Onoue, Kentarou Kurashige, “A relationship between ability of perception and learning efficiency”, Proceedings of World Automation Congress 2008, 2008.

研究業績

[1] Kentarou Kurashige, Yukiko Onoue, “The robot learning by using sense of pain”, Proceedings of International Symposium of Humanized Systems 2007, pp.1-4, 2007.

[2] Yukiko Onoue, Kentarou Kurashige, “A relationship between ability of perception and learning efficiency”, Proceedings of World Automation Congress 2008, 2008.