目次

第1章　序論

1.1　ロボットが使用される環境

近年，ロボットが周囲の環境に合わせて自律的に行動を学習する研究およびその実用化が進んでいる．実用化され始めた頃のロボットは，人間によって事前に設定された動作を繰り返すものがほとんどであり，主に工場での組み立て作業などに用いられていた[1]．そのためロボットが使用される環境は限定された場所であり，ロボットにとって最適化されていた．工場の組み立てラインで作業するロボットを例に挙げると，工場のラインという限定された環境で，人間によって設定された組み立てという作業を延々と繰り返し行うのみである．現在でも組み立てロボットのような産業用ロボットは使用されており，これらはロボットが使用され始めた頃に比べて，移動機能やセンシング機能などが向上した．その結果同じ作業を繰り返すだけのロボットから，状況に応じた行動を取ることが可能となった．

現在ロボットが活躍する場は，工場などの作業を行う場だけでなく，自然環境や家庭環境，オフィスなど人間の生活に近い場所へと広がっている．人間の生活環境は，実用初期のロボットが使用されていたような限定かつ最適化された環境に比べ複雑であり，環境は時間によって変化する．家の中を例に挙げると，配置されている家具や雑貨などの位置は日々変化する可能性があり，一度ある場所に存在していたものがいつまでもそのままの場所にあるとは限らない．ソファの上にあった雑誌が翌日には床の上に落ちている可能性もあるし，テーブルの上にあった箱ティッシュがいつのまにか捨てられていることもある．このように，ロボットの進化に伴って使用される環境が図 1.1 のように変化している．

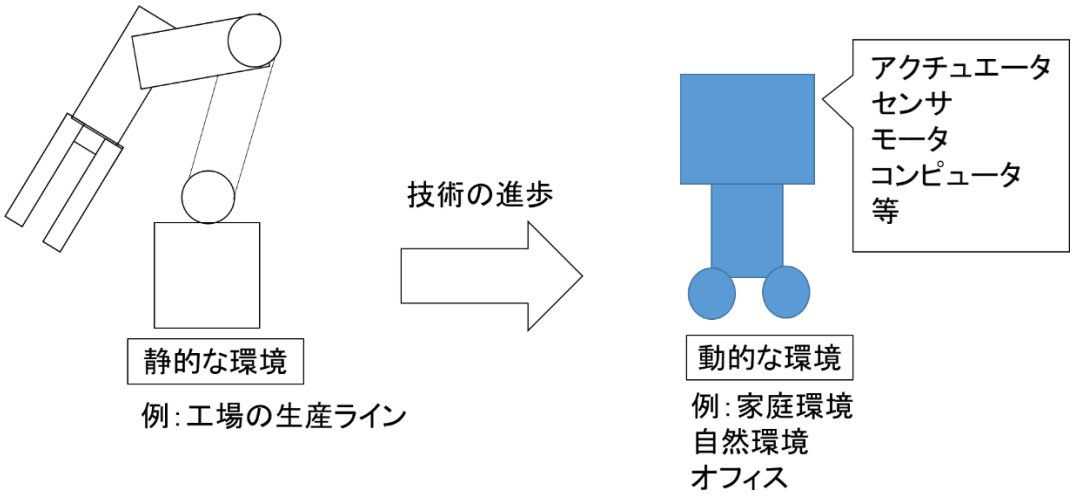

図 1.1　ロボットの進化と使用される環境の変化

1.2 学習機能によるロボットの環境への適応

ロボット実用初期のような静的な環境においては，環境の変化がほとんどないため，ロボットが直面する環境やロボットの状況を人間が事前に予測することが可能であった．そのため，環境に対するロボットの行動を人間が事前に設定することができていた．しかし，家庭環境など変化が多い複雑で多様な環境下においては，ロボットが直面し得る環境全てを人間が予測することはできない．そこで，複雑かつ多様な環境下では，ロボットが自律的に周りの環境に応じた行動を取ることが求められる．それを実現するための方法の一つとして，ロボット自身に学習機能を持たせる方法がある[2-3].

ロボットが学習機能を持つということは，ロボットが経験によって得た知識を基に行動を行うことである．ロボットに行動を選択するための規則やルール，判断基準などを与え，それらに従うことでロボットは環境に応じた行動を自律的に学習することができる．よってロボットは未知の環境下であってもその環境に応じた行動を取ることが可能となる．

このようにロボットは，学習機能を得ることによって変化していく環境にある程度適応することが可能となった．

1.3 ロボットの学習方法

ロボットが未知の環境において学習する方法として，強化学習というものが存在する[4]. 強化学習とは，ロボット等の学習する対象が環境に対して行試行錯誤することで，環境に応じた最適な行動を自律的に学習する方法である．

強化学習では，ロボット等の学習する対象が正しい行動を教えられるのではなく，行動を決めるルールに従って行動を行い，その行動を評価することで学習を行う．具体的には，ロボットが自身の現在の状態を認識し，ルールに従って状態に応じた行動を行い，その結果に対して報酬を与えることで価値関数を作成する．報酬は，ロボットに対して人間が設定する．そして価値関数を基に最適なルールを作成し，よりよい行動を選択していく．これらの手順を繰り返すことで，最適な行動を学習する．詳しい説明は第 2 章で行う．本研究では，この強化学習を適用したロボットを対象としている．

1.4 強化学習を適用したロボットの実用例

強化学習を適用したロボットの例として，経路探索問題における行動学習について述べる[5]．この場合の経路探索のような仕事のことをタスクと呼ぶ．このタスクの内容は，図 1.2 のように，ロボットがスタート地点を出発してゴール地点へ到達してタスクを達成するまでの経路を通る行動を学習するものである．ロボットはゴール到達時点で報酬を与えられ，報酬を得るために試行錯誤を繰り返す．この時ロボットは，スタートからゴールまでの最短経路を学習する．

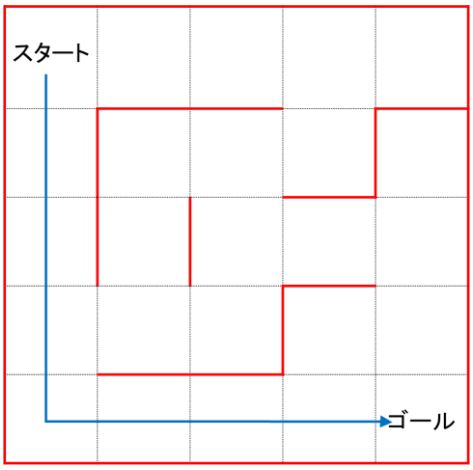

図 1.2 経路探索問題

1.5 タスク達成のためのサブゴール

1.3 節では，強化学習を適用したロボットがタスクを実行し，達成するための行動を学習する例として経路探索問題について述べた．経路探索問題において，図 1.3 のようにある一つの経路を学習させたい時，ゴールまでの経路の途中に小目標を設定し，小目標をクリアするように行動を学習させる．この小目標をサブゴールと定義する．この場合，サブゴールは学習させたい経路によって設定する場所が異なる．

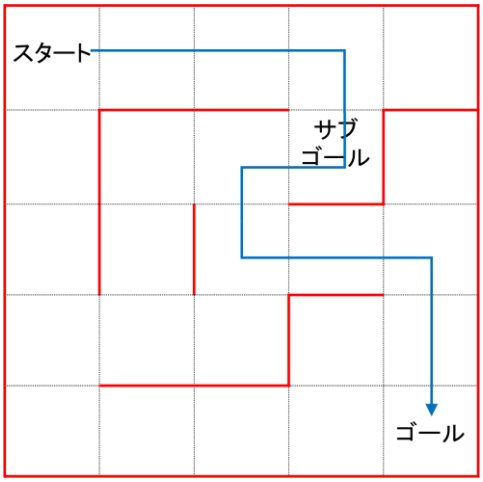

図 03 サブゴールの存在するタスク

1.6 従来研究

1.6.1 サブゴールを扱う従来研究

サブゴールを扱う従来研究として，ロボットにサブゴールをクリアする行動を学習させる研究が存在する．“Shaping 強化学習を用いた自律エージェントの行動獲得支援手法” [6]，“遅れ報酬に基づく遺伝的アルゴリズムによる部分観測マルコフ決定問題の解決手法” [7]，“複数の学習器の階層的構築による行動獲得” [8]では，いずれもサブゴールをクリアすることでロボットに報酬を与えることで，サブゴールをクリアする行動の学習を可能としている．

また“移動ロボットによるサブゴール間巡回行動の学習” [9]という研究においては，あらかじめ設定されたサブゴール間の効率的な移動をロボットに学習させることを目的としている．この研究では，サブゴールをロボットがその上に来た場合のみ認識できる場所と定義しており，認識の方法としては，ロボットに信号を送ることで行っている．ロボットはまずサブゴールの位置を探索するために移動を行い，発見されたサブゴール間を最も効率的に移動できる行動を学習する．このサブゴール間の移動経路の学習は，移動ロボットが障害物を回避する経路の学習のみを目標としており，障害物回避後の目標位置への到達については学習することができない．

1.6.2 従来研究における問題点

タスクを達成するための条件は様々存在し，その中でもタスクを達成するための小目標（サブゴール）が存在すると述べた．この時サブゴールはタスクや学習環境などによって異なる．経路探索問題を例にすると，図 1.4 のようにロボットに学習させたい経路や壁の位置によって，設定するサブゴールが異なる．

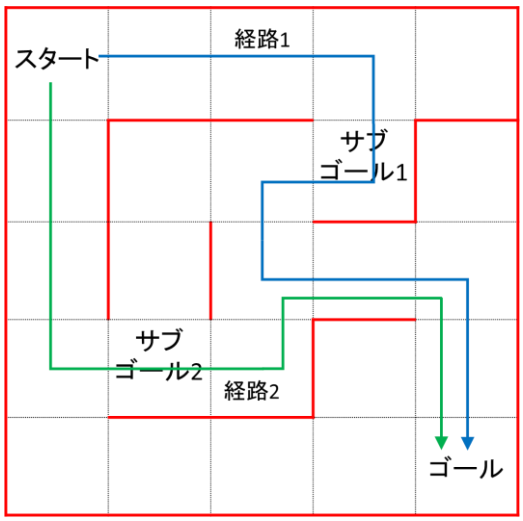

図 1.4 学習経路や環境によってことなるサブゴール

また図 1.5 のように壁の位置などが事前に想定することができない場合，ロボットが取り得る経路や壁の位置などが想定できず，サブゴールの設定が困難である．

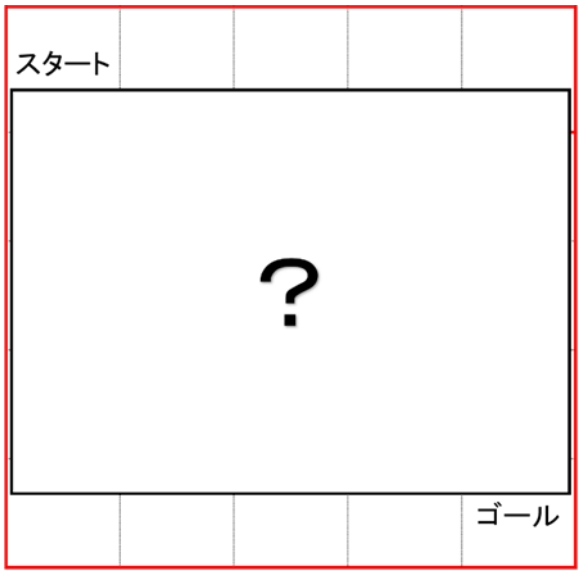

図 1.5 環境の想定が困難な場合

従来研究でのロボットに適用されている学習方法は，ロボットがサブゴールをクリアした時に報酬を与えられる，もしくは人間から知らされることで，サブゴールを認識し，サブゴールをクリアする行動の学習を可能としている．前者の場合，与える報酬は人間が設定するため，人間が環境を事前に想定する必要がある．後者の場合，人間がロボットにサブゴールのクリアを知らせるために，ロボットの状態を常に認識している必要がある．

強化学習を適用したロボットは，行動を自律的に学習するという特性上，未知の環境下で利用されることが多い．その未知の環境下においては，環境が複雑に変化する可能性があるため，ロボットが直面する環境や状況を事前に想定することは難しい．その環境の中ではあらかじめサブゴールを想定することも困難であり，ロボットに対してサブゴールでの報酬設定を行う方法では，あらゆるサブゴールを想定して報酬を設定する必要があるか，もしくは設定することができない．またロボットにサブゴールのクリアを知らせる方法では，自律的に学習を行うという強化学習の特徴に反しているので好ましくない上，常にロボットの状態を認識していなければならない．

1.7 本研究の目的

本研究では，あらゆる環境においてロボットがサブゴールをクリアする行動の学習を実現するために，サブゴールを自律的に発見しクリアする行動学習のできるロボットを目標とする．従来研究では人間の手によってロボットにサブゴールを認識させていたが，事前に想定することが困難でかつ動的に変化する環境下においては，タスクに適したサブゴール

の設定を人間が行うことに限界がある．そのためロボットが人間の手を借りず自律的にサブゴールを発見し，サブゴールをクリアする行動を学習する必要がある．

そこで本研究では，ロボットが自力でサブゴールを発見するための手掛かりとして，ロボットの過去の経験を用いる．そして過去の経験を基にサブゴールを発見し，見つけたサブゴールをクリアする行動を学習する学習システムを提案する．提案する学習システムによって，サブゴールをクリアしてタスクを達成する行動を学習するロボットを実現する．ロボットの学習には，最も一般的な学習手法である強化学習を用いる．

1.8 サブゴール発見の方針

本節では，ロボットがどのようにして自律的にサブゴールを発見するか方針を立てる．ロボットが学習を行う環境が事前に想定できず，サブゴールでの報酬設定が困難な場合であっても，タスクを達成するためにゴールでの報酬設定は可能である．そこでサブゴールをクリアしているか否かで，ゴールで与える報酬の大きさに差をつけることを考える．

本研究で用いる強化学習は，動物のしつけに近い方法で行動を学習させる[10]．強化学習において，褒める行為は，報酬を与えることと同じである．これを利用して，ロボットがサブゴールをクリアしてタスクを達成した場合，クリアしなかった場合よりもさらに大きい報酬を与えることで，サブゴールをクリアする行動が求められていることを学ばせる．このイメージ図を図 1.6 に示す．

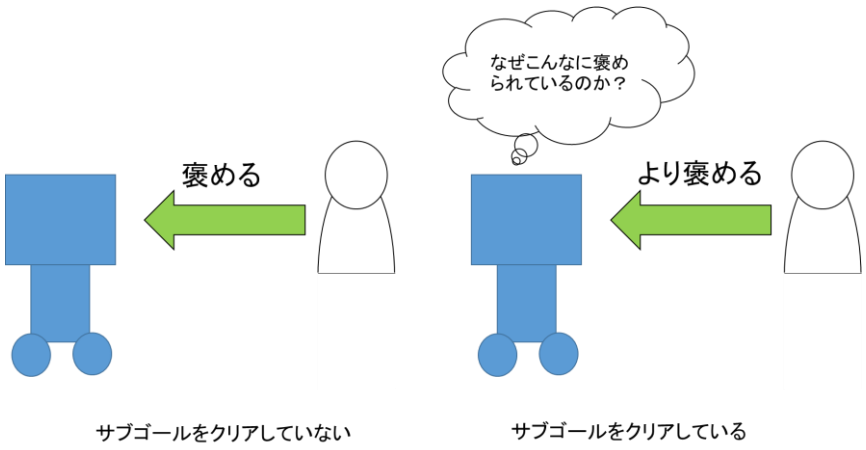

図 1.6 ロボットがサブゴールを発見する方針

しかしただ大きい報酬を与えるだけでは，その原因をロボットは理解していない．そこでロボットに過去の経験を蓄積させ，何を経験したことでサブゴールをクリアしなかった時よりも大きい報酬を得られたのかを求める．

サブゴールをクリアする行動を学習するための手段として，ロボットが環境から与えられる状態の他に，サブゴールをクリアしているか否かという状態を追加する．これによって，環境から状態を与えられた時，サブゴールをクリアしている場合としていない

場合で別の状態となる．

1.9 本論文の構成

第 1 章では， 学習機能をもつロボットが使用される環境の変化と，サブゴールについて述べた．

第 2 章では，本研究で対象とする学習手法である強化学習について説明し，ロボットに適用した場合どのような流れで学習を行うのかを説明する．

第 3 章では，提案する学習システムについて説明する．初めに強化学習においてサブゴールをどのように扱うかを述べ，その後システムの概要，構成を述べる．終わりに，提案するシステムにおける学習の方法について詳しく述べる．

第 4 章では，本研究で提案するシステムの検証実験について述べる．また実験によって得られた結果について考察する．

第 5 章では，検証実験の結果を基に全体のまとめを述べ，今後の課題について説明する．

第 2 章　強化学習

2.1 強化学習の概要

強化学習[5]とは，環境に対して試行錯誤的に行動を選択することで，適切な行動を獲得する機械学習の一種である．ロボットに強化学習を適用することで自律的に学習を行うことが可能となる．その結果周りの環境に適応し，より良い行動を選択することが可能となる．

強化学習では，報酬というスカラー値を用いて学習を行う．ロボットと環境の関係は図 2.1 のようになっており，ロボットが周囲の環境の状態に対して行動し，その行動に応じて環境から報酬を獲得することができる．ロボットは環境に対する試行錯誤を通して，報酬が最も多く得られる行動を選択する．第 1 章で強化学習が動物のしつけに近いと述べたのは，このためである．報酬は人間が事前に設定する必要があり，ロボットが行うタスクに応じて報酬の設定をする．

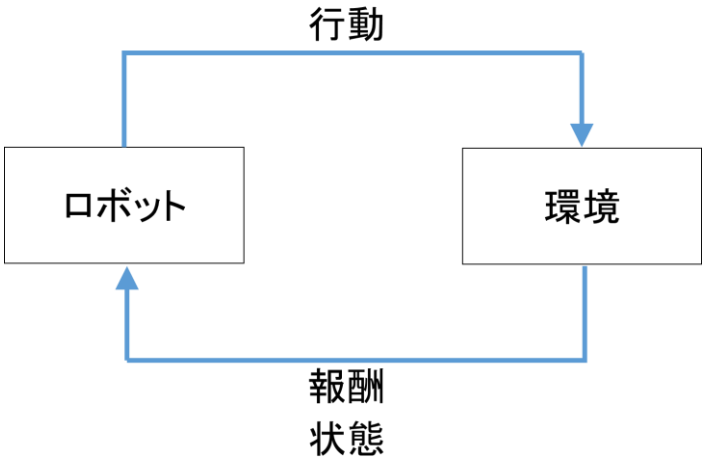

図 2.1 ロボットと環境の関係

強化学習では，タスクに対して何が最適な行動なのかをロボットが考慮する必要はない．人間がタスクに対してロボットに行って欲しい行動ほど高い報酬を与え，好ましくない行動に対しては低い報酬を与えることで，ロボットは自動的に最善の行動を獲得することができる．これによって強化学習を適用したロボットは未知の環境を扱うことが可能であり，実ロボットによく用いられる．

2.2 強化学習における学習の流れと学習方法

前節では，強化学習の概要について述べた．本節では強化学習における学習の流れを説明し．具体的な学習方法について述べていく．

強化学習で対象とするロボットには何らかのセンサが搭載されており，センサを通じて周囲の状態を認識することが可能である．また，ロボットは認識した状態に対して何らかの行動を取ることができる．このようなロボットに対する強化学習の概念図を図 2.2 に示す．

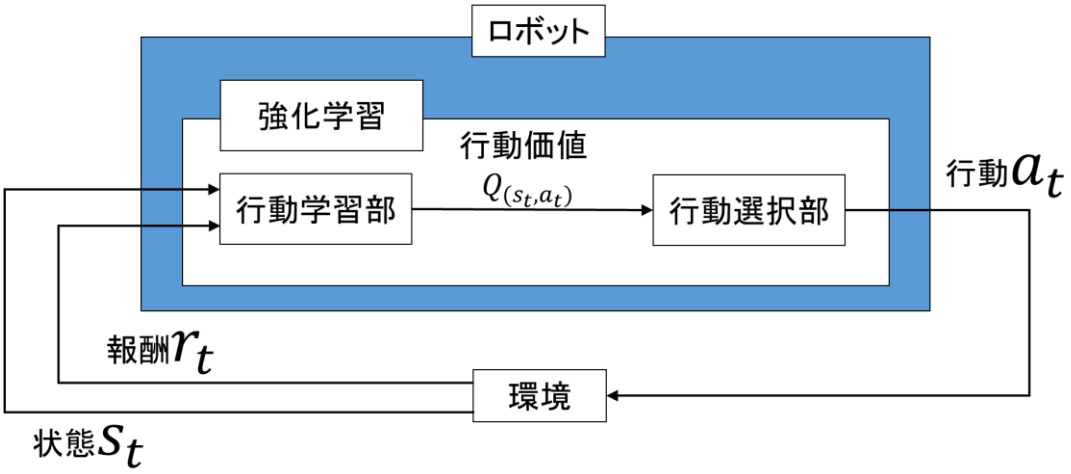

図 2.2 強化学習の概念図

時刻 t においてロボットが認識した状態をs_tとする．ロボットは認識した状態s_tに対して，過去の学習結果から行動a_tを選択する．この時選択した行動に対して環境から報酬r_tを獲得し，獲得した報酬r_tと状態s_tは行動学習部に入力される．行動学習部では状態s_tと行動a_tの組み合わせに対する行動価値を更新する．この状態s_t行動a_tの組み合わせを状態行動対と呼ぶ．行動価値とは，ロボットが取った行動によって獲得することのできる報酬の期待値である．報酬がロボットの取った行動に対して即時的な意味合いで良し悪しを示すものであるのに対し，行動価値は将来的な行動の望ましさを示している．行動価値の更新方法は，用いる行動学習法によって異なるので，詳しくは 2.4.節で説明する．行動学習部で算出された行動価値は行動選択部へ入力される．行動選択部では，行動学習部で更新された行動価値から次の行動を選択する．行動を選択する方法は，用いる行動選択手法によって異なるので，詳しくは 2.5 節で説明する．ロボットは行動学習と行動選択を繰り返すことで，目的遂行のために最適な行動を学習することが可能となる．これまでの流れを図 2.3 に示す．

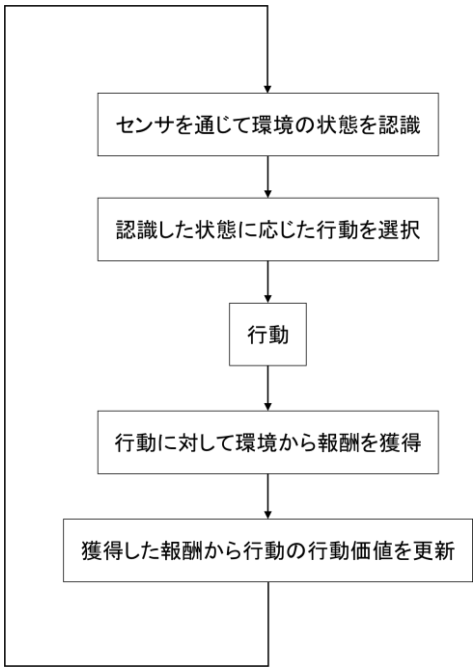

図 2.3 強化学習の流れ

2.3 ロボットが行動する環境モデル

強化学習においては多くの場合，ロボットが行動する環境をマルコフ決定過程（Markov Decision Process，MDP）によりモデル化する．MDP とは次のようなモデル化を指す．環境の取り得る状態の集合を$\boldsymbol{S} = \{s_1, s_2, \cdots, s_n\}$，ロボットが取り得る行動の集合を$\boldsymbol{A} = \{a_1, a_2, \cdots, a_n\}$と表した時，環境中のある状態$s \in \boldsymbol{S}$において，ロボットがある行動$a \in A$を実行すると，環境は確率的に状態$s' \in \boldsymbol{S}$へ遷移する．この時環境からロボットへ報酬$r(s,a)$が確率的に与えられる．この流れを図 2.4 に示す．状態遷移と報酬は現在の状態と行動のみに依存し，それ以前の状態や行動の履歴には依存しない．

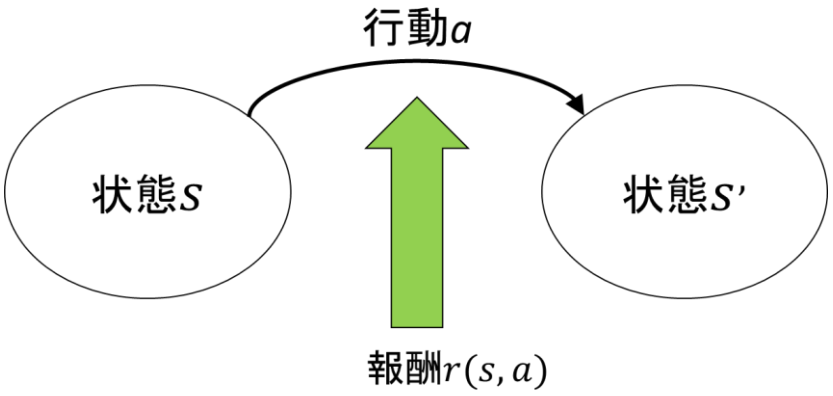

図 2.4 MDP による環境のモデル化

MDP 環境下における強化学習では，ロボットは環境の状態を完全に認識できると仮定している．またロボットは 報酬の与えられ方についての知識をあらかじめ持たないとされており，そのため学習の際には将来に獲得できる報酬を予測する必要がある．

またロボットは報酬の与えられ方をあらかじめ知らされていないので，報酬を与えられて初めてその状態で報酬が得られることがわかる．

2.4 行動学習手法

行動学習手法は 2.2 節で説明した行動学習部において，環境から獲得した報酬から行動価値を評価する手法である．

2.4.1 加重平均法

加重平均法は，遠い過去に受け取った報酬を考慮するのか，または近い時刻に受け取った報酬のみを考慮するのかを重み付けによって変更し，行動価値を更新する方法である．ある時刻 t において状態 s で行動 a を取った場合の行動価値$Q_t(s,a)$の更新式を式(2.1)に示す．αは学習率と呼ばれる定数で，この値を変化させることで与える重みが変化する．αの値を大きく設定した場合には，近い時刻に受け取った報酬の影響が大きくなり，αの値を小さく設定した場合には，遠い過去に受け取った報酬も考慮した行動価値の更新となる．

$$Q(s_t,a_t) = Q(s_t,a_t) + \alpha[r_{t+1} - Q(s_t,a_t)] \qquad (2.1)$$

r_{t+1}：新たに獲得した報酬

2.4.2 Q 学習

Q 学習は，現在の状態で選択した行動の価値と，その行動の結果遷移した先の行動価値によって，現在の行動価値を更新する方法である．ある時刻 t において状態s_tで行動a_tを取った場合の行動価値$Q_{(s_t,a_t)}$の更新式を式(1)に示す．ここでαは学習率，γは割引率と呼び，いずれも(0,1]の定数である．αの値を大きく設定すると近い時刻に受け取った報酬の影響が大きくなり，値を小さく設定すると遠い過去に受け取った報酬も考慮した行動価値の更新となる．γの値を大きく設定した場合は将来の行動価値を重視し，小さく設定場合は現在の行動価値を重視する．

$$Q_{(s_t,a_t)} \leftarrow Q_{(s_t,a_t)} + \alpha\left[r_{t+1} + \gamma \max_a Q_{(s_{t+1},a)} - Q_{(s_t,a_t)}\right] \qquad (2.2)$$

$Q_{(s_t,a_t)}$：時刻 t，状態s_tで行動a_tを選択した場合の行動価値

r_{t+1}：新たに獲得した報酬

2.5 行動選択手法

行動選択手法は 2.2 節で説明した行動選択部において，行動価値から最適な行動を選択するための方法である．ここでは，代表的な手法としてε-greedy法とsoftmax法について説明する．

2.5.1 greedy法

greedy法は，ある状態において最も高い行動価値を持つ行動を選択する．この方法は，常に行動価値が最大の行動を選択していくため，一時的に行動価値が低いだけで，本当は価値の高い行動である可能性を確かめることを行わない．

2.5.2 ε-greedy法

ε-greedy法は，過去の行動価値の中からεの確率でランダムな行動を選択し，$1-\varepsilon$の確率で行動価値の最も高い行動を選択する手法である．εの確率でランダムな行動を取ることで，現在最も高い行動価値となっている行動よりも更に高い行動価値を持つ行動があるかどうかを探索することができる．

2.5.3 softmax法

softmax法は行動価値によって行動を選択する確率を変化させる方法である．この確率のことは行動確率と呼ぶ．行動価値の高い行動には最も高い行動確率が与えられ，その他の行動には，行動価値の高い順に行動確率が与えられる．時間 t において状態s_tにおいて行動a_tを選択する確率$\pi_t(s,a)$は式(2)で与えられる．$Q_{(s,a)}$はある時刻 t において状態sで行動aを取った場合の行動価値で，τは温度と呼ばれる正の定数である．

$$\pi_t(s,a) = \frac{e^{Q_{(s,a)}/\tau}}{\sum_{b=1}^{n} e^{Q_{(s,b)}/\tau}} \qquad (2.3)$$

温度τが高い場合には，すべての行動がほぼ同程度に選択され，低い場合には行動価値高低による選択確率の差が大きくなる．

2.6 強化学習における問題点

強化学習では，ある状態において行動価値の最も高い行動を学習する．もし一つの状態に対して最大の行動価値を持つ行動が複数存在する場合，行動を一つに決めることができずランダムに行動を選択してしまう．

第 3 章　提案システム

本章では，ロボットがサブゴールの存在を知らされていない状況下において，過去の経験と獲得した報酬の差を基にサブゴールの発見を行い，発見したサブゴールをクリアしてタスクを達成する行動学習を行う学習システムを提案する．本論分で想定するサブゴールは 1 つで，消滅や増加，変化はしない．第 1 節では強化学習におけるサブゴールの定義について述べ，第 2 節以降で提案システムの具体的な構造や学習方法などについて述べる．

3.1 強化学習におけるサブゴール

本研究では図 3.1 のように，ロボットがゴールへ到達した時点で報酬を与えられ，ある特定の状態を経験することで報酬の大きさが異なる場合，この特定の状態をサブゴールとする．人間はロボットがサブゴールをクリアしてもその時点で報酬は与えず，信号を送るなどの方法で存在を知らせることもしない．よって従来研究のように，人間の手によってロボットにサブゴールの存在を知らせるような方法では，サブゴールをクリアしてタスクを達成する行動を学習することはできない．

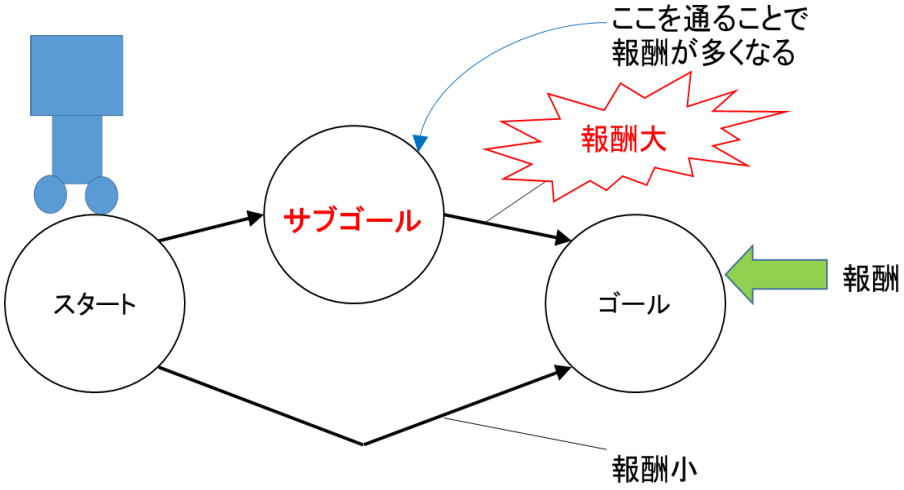

図 3.1 強化学習におけるサブゴールの扱い

3.2 提案システムの概要

本論文で提案するシステムでは，まず学習機能を持つロボットが経験し得る全状態の内，どの状態がサブゴールにあたるかを探索する．この時ロボットは行動学習を行わない．探索終了後は行動学習に移る．学習を行う際は，探索の結果発見したサブゴールをクリアして，ゴールへ到達する行動を学習する．探索と行動学習を分けて行うのは，行動

学習が探索の妨げになるおそれがあるためである．

3.3 提案システムの構成

本システムは，サブゴールの発見を行う部位と，サブゴールをクリアする行動の学習を行う部位で構成される．サブゴールの発見はサブゴール探索部で行い，行動の学習はサブゴールクリア判定部と強化学習を行う部位によって行う．

サブゴール探索部では，ロボットが認識した現在の状態を蓄積し，取り得る行動の中からランダムに行動を選択する．またタスクの開始から達成までを 1 試行とすると，各試行終了時，ゴールで獲得した報酬の大小によって経験した状態を分類し蓄積する．これらの動作を特定の試行まで行った後，蓄積された状態からサブゴールに当たる状態を割り出し，その結果発見したサブゴールを，サブゴールクリア判定部へ渡す．

サブゴールクリア判定部では，渡されたサブゴールをクリアしているか否かを判定する．発見したサブゴールをクリアしたと判定した後，ロボットが認識する状態を追加し，状態行動対を多くする．行動の学習は強化学習によって行い，ゴール到達時に報酬を与える．

3.4 提案システムの動作

3.4.1 サブゴールの探索と発見

まずサブゴールを発見するために，サブゴールを探索する．探索時における提案システムの動作を図 3.2 に示す．

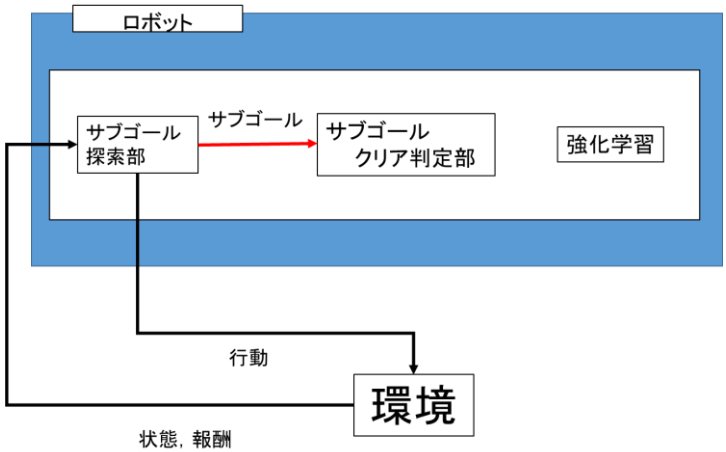

図 3.2 提案システムの動作：サブゴール探索時

サブゴール探索部では，1 試行中にロボットが認識し，経験した状態を蓄積する．その後自身が取り得る行動の中からランダムに行動を選択し，行動の結果に応じて報酬r_tを与えられる．試行終了時，獲得したr_tの値に応じて，試行中に経験した状態に分類し

蓄積する．ある試行 t において与えられる報酬r_tを式(3.1)のように定義する．

$$r_t := \begin{cases} r_{small}(\text{サブゴールをクリアしていない}) \\ r_{big}(\text{サブゴールをクリアしている}) \end{cases} \left(r_{small} < r_{big}\right) \quad (3.1)$$

この時，ロボットが t 試行目に経験した状態の集合を$\boldsymbol{E}_t$とすると，$r_t = r_{big}$の時常に経験した状態の集合$\boldsymbol{E}_{big}$と，$r_t = r_{small}$の時一度でも経験したことのある状態の集合$\boldsymbol{E}_{small}$にそれぞれ蓄積する．$\boldsymbol{E}_{big}$，$\boldsymbol{E}_{small}$それぞれを式(3.4)，式(3.5)で定義する．

$$\boldsymbol{E}_{big} := \bigcap_{t \in T_{big}} \boldsymbol{E}_t \quad (3.2)$$

$$\boldsymbol{E}_{small} := \bigcup_{t \in T_{small}} \boldsymbol{E}_t \quad (3.3)$$

$$T_{big} := \{t | t \in \boldsymbol{N} \wedge r_t = r_{big}\}$$

$$T_{small} := \{t | t \in \boldsymbol{N} \wedge r_t = r_{small}\}$$

$\boldsymbol{N}$：自然数の集合

特定の試行終了後サブゴールの探索を終了し，$\boldsymbol{E}_{big}$および$\boldsymbol{E}_{small}$からサブゴール sG を割り出す．サブゴールとなり得る状態は，報酬が小さい試行では一度も経験しておらず，報酬が大きい試行では必ず経験している状態である．このことから，サブゴール sG を式(3.6)のように求め，発見したサブゴールをサブゴールクリア判定部へ渡す．また式(3.6)をベン図で表したものを図 3.3 に示す．

$$\text{sG} := \boldsymbol{E}_{big} \cap \overline{\boldsymbol{E}_{small}} \quad (3.4)$$

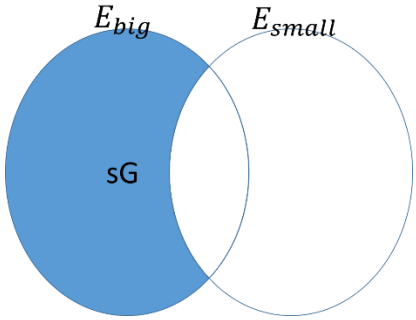

図 3.3 式(3.6)を表したベン図

3.4.2 サブゴールをクリアする行動の学習

サブゴールを発見した後は，発見したサブゴールをクリアしてからタスクを達成する行動を学習する．この時の提案システムの動作を図 3.4 に示す．

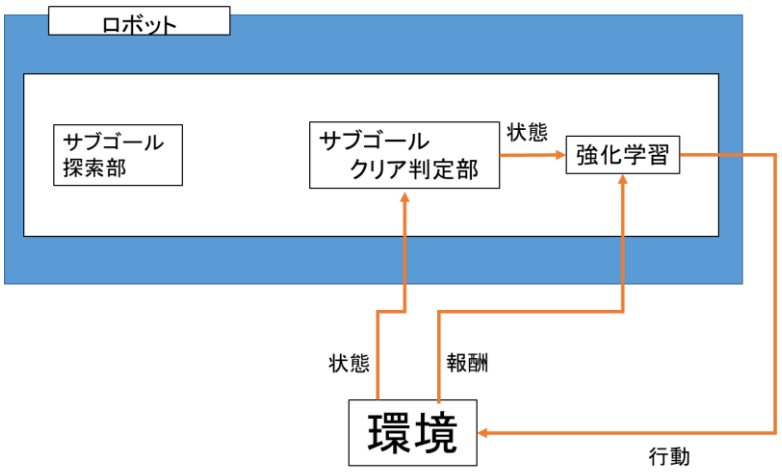

図 3.4 提案システムの動作：行動学習時

まずロボットは現在の自身の状態を認識する．次に認識した状態が発見したサブゴールと一致するか否かを，サブゴールクリア判定部にて判定する．その後サブゴール判定部から状態が出力され，その状態を認識した状態として強化学習により行動学習を行う．

サブゴール判定部にはサブゴール探索部から渡されたサブゴールが蓄積されており，ロボットが認識した状態が渡されたサブゴールと一致していれば，サブゴールをクリアしたことにする．判定の結果サブゴールをクリアしている場合，図 3.5 のように状態を追加する．一度サブゴールをクリアしたと判定した後は，1 試行中常に状態を追加したままである．

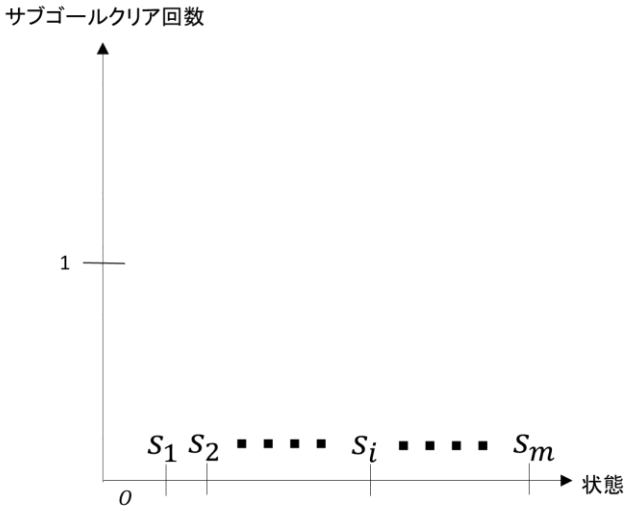

図 3.5 発見したサブゴールのクリアに伴う状態追加

状態を追加する理由としては，サブゴールをクリアするまでの行動と，クリアした後の行動を別々に学習するためである．強化学習では，ある状態行動対における行動価値の値が最大となる行動を選択する．しかし一つの状態行動対において同値となる行動価値が複数存在する場合，学習する行動を一つに決めることができない．よって一つの状態において，サブゴールをクリアするまでの行動と，サブゴールをクリアした後の行動の両方を学習する必要がある．

第 4 章 実験

4.1 実験の目的

本実験はシミュレーション実験で行う. 本実験の目的は本論文で提案するシステムの有用性を検証することである. 具体的には, 提案するシステムを適用したロボットがサブゴールを発見し, 発見したサブゴールをクリアする行動が学習できるかどうか検証する.

4.2 実験概要

本実験では, サブゴールが存在する経路探索問題を扱う. ロボットは強化学習によってスタートからゴールまでの経路を学習する.

4.3 実験設定

本実験の実験環境は図 4.1 のような 3×3 のグリッドワールドで, 周囲と内部に壁が存在する. スタート, ゴール, サブゴールが存在し, ロボットはゴールに到達すると報酬を得られる. またサブゴールを通過してゴールへ到達すると, より多くの報酬を得ることができる. サブゴールの数は 1 つで位置は固定, ロボットはサブゴールの位置と存在を知らない. スタートからゴールへ到達するまでを 1 試行とし, ゴールへ到達するとスタート位置まで戻る.

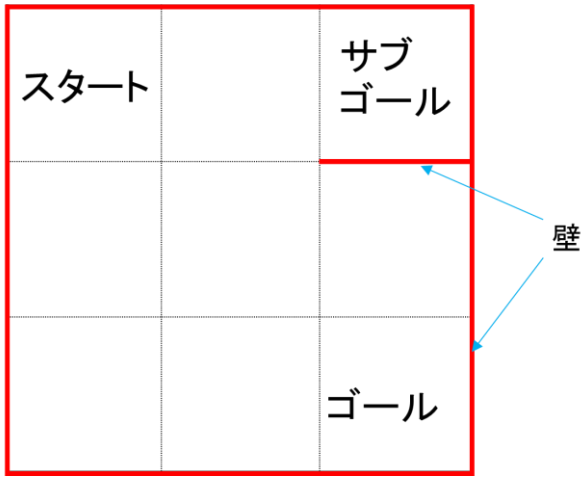

図 4.1 実験環境

今回は目的を達成するために, 二体のロボットを想定する. これらのロボットをそれぞれロボット A, ロボット B と呼ぶことにする. ロボット A は強化学習によってスタートからゴールまでの行動を学習するロボットで, ロボット B は提案システムを適用した

ロボットである．

ロボットBは，自らが通過した場所を蓄積する機能を持ち，通過した場所を試行毎に蓄積する．他にも，ゴールで得られる報酬量が大きかった時常に通過している場所，報酬量が小さかった時に一度でも通っている場所それぞれを蓄積する．また，特定の試行まではサブゴールを探索するために動く．この間ロボットは全てランダム行動を行い，行動学習を行わない．探索が終了した時点でサブゴールを発見する処理を行い，その後残りの試行で発見したサブゴールを通過してゴールへ到達する行動を学習する．特定の試行までのロボットの全状態は図 4.2 のようになっており，サブゴールの探索終了以降は状態数を追加し，図 4.3 のようになる．

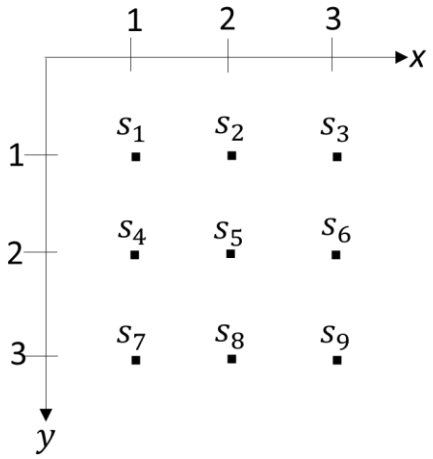

図 4.2 ロボットが経験し得る全状態

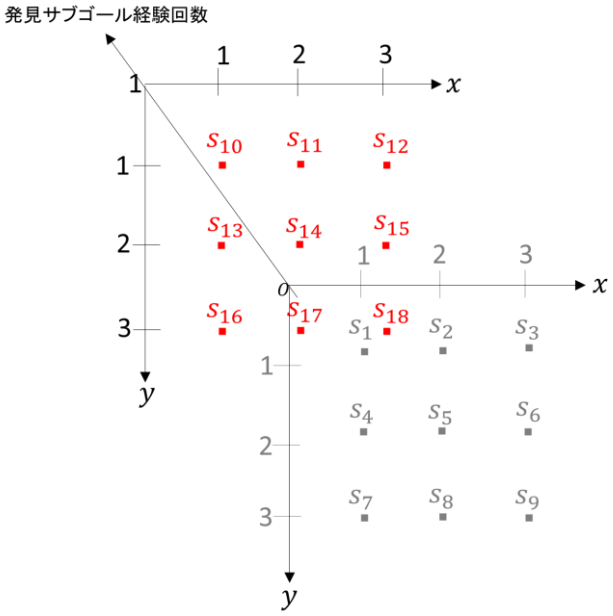

図 4.3 サブゴール経験に伴う状態追加

各ロボットが取ることのできる行動は共通で，上下左右へ移動が可能であり，壁に衝突するとその場に停止する．行動学習手法は Q 学習を，行動選択手法はε-greedy法を使用する．

本実験のパラメータを表 1 に表す．

表 1　パラメータ

学習率	0.10
割引率	0.90
報酬	100.0（サブゴール未通過） 1000.0（サブゴール通過）
試行回数	20000 回(行動学習時) 1000 回(サブゴール探索時)
ε (ε − greedy 法)	0.05
サブゴールの数	1

4.4 実験結果

シミュレーション実験の結果を示す．　まず各ロボットの行動学習時における試行毎にゴールで獲得した報酬の一例を図 4.4，図 4.5 に示す．サブゴール探索時は行動学習を行っていないので, 与えられた報酬から考察できることはないと判断したため示さないこととする．

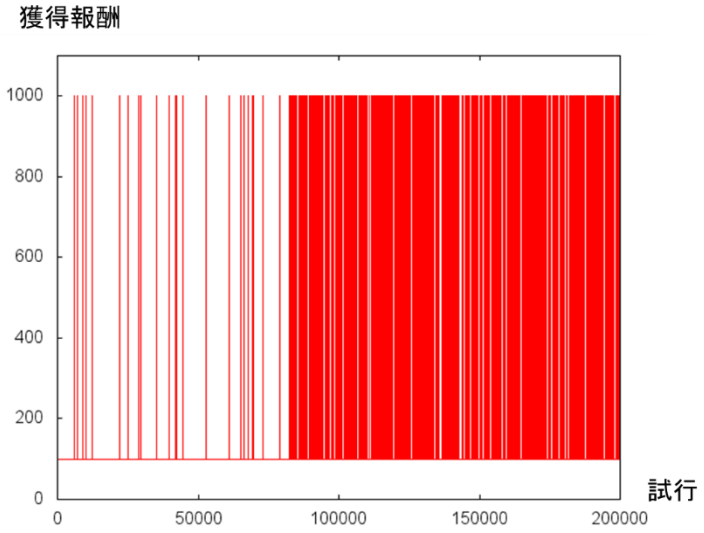

図 4.4 ロボット A が試行毎にゴールで獲得した報酬

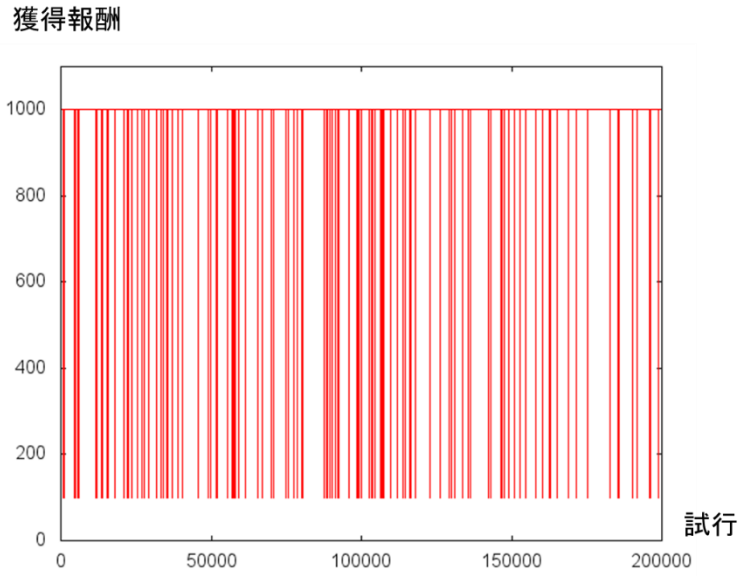

図 4.5 ロボットＢが試行毎に獲得した報酬

次に，前半 10 万試行で与えられた報酬を図 4.6，図 4.7 に示す.

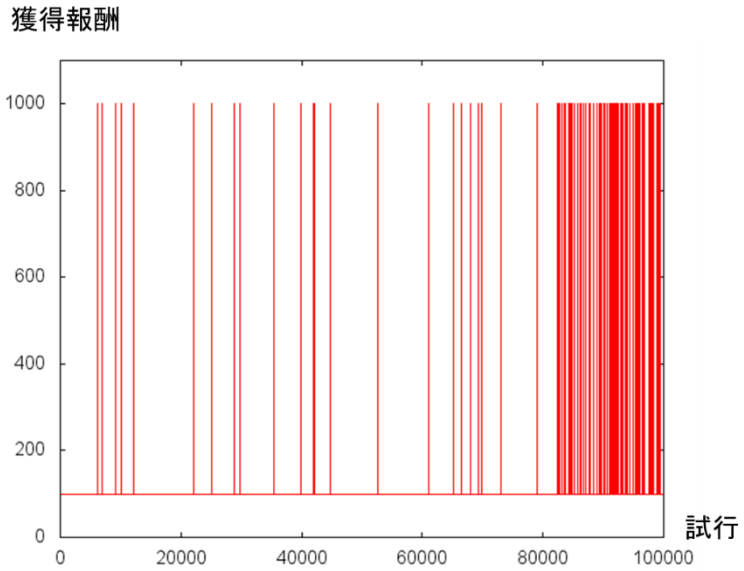

図 4.6 ロボットＡが獲得した報酬(0～100000 試行)

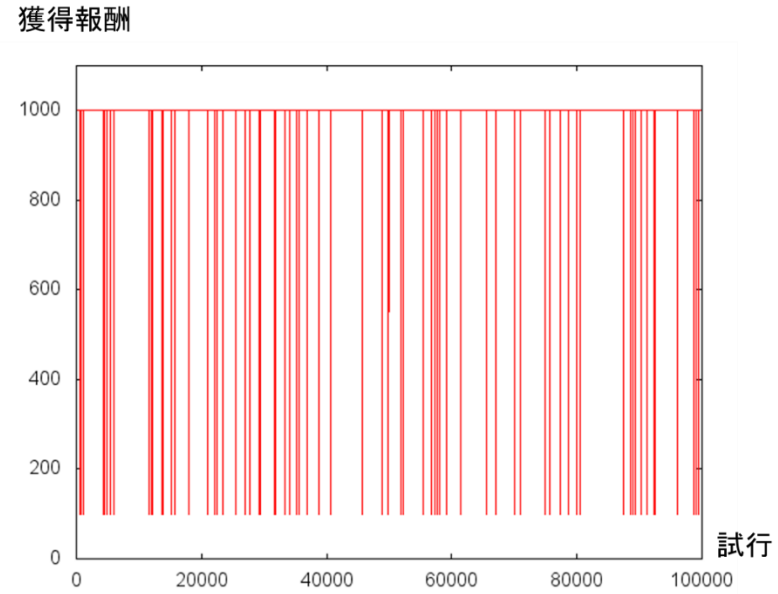

図 4.7 ロボットＢが獲得した報酬(0～100000 試行)

さらに，最後 10 万試行で与えられた報酬を図 4.8，図 4.9 に示す.

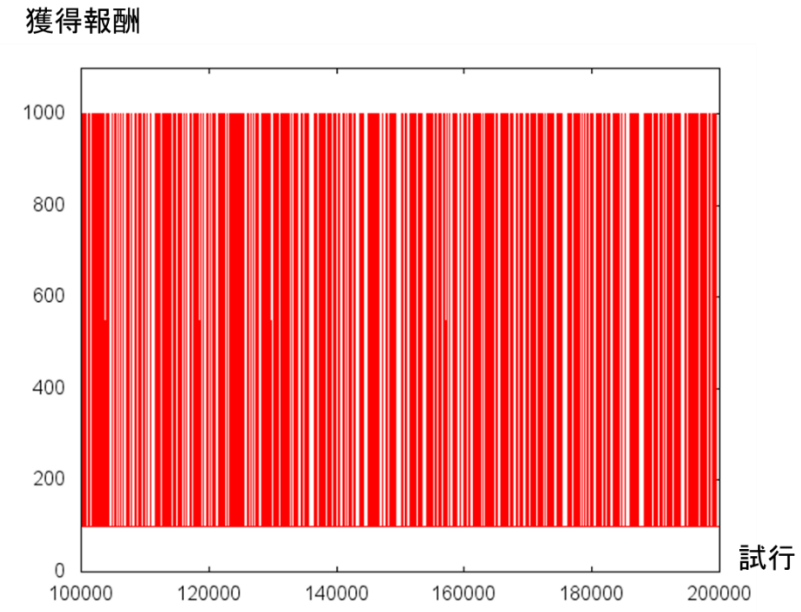

図 4.8 ロボットＡが獲得した報酬（10000～200000 試行）

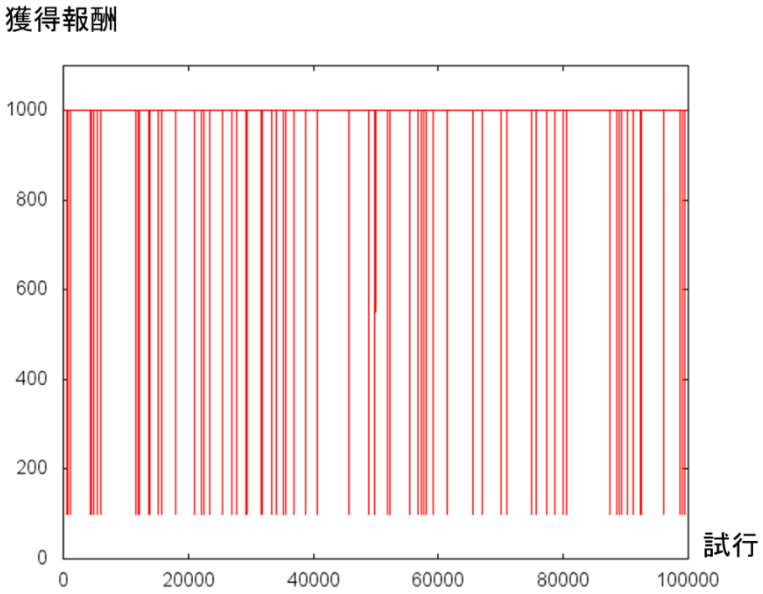

図 4.9 ロボットＢが獲得した報酬(100000～200000 試行)

始め 10000 試行で与えられた報酬を図 4.10，図 4.11 に示す.

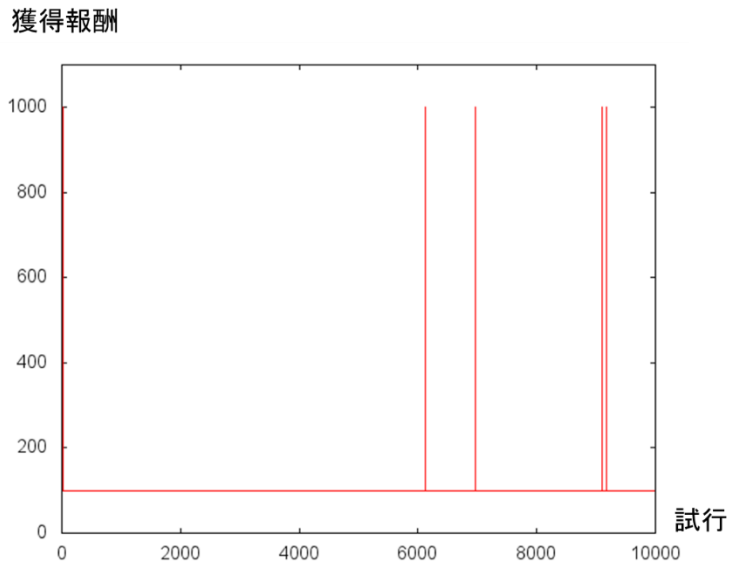

図 4.10 ロボットＡが獲得した報酬(0～10000 試行)

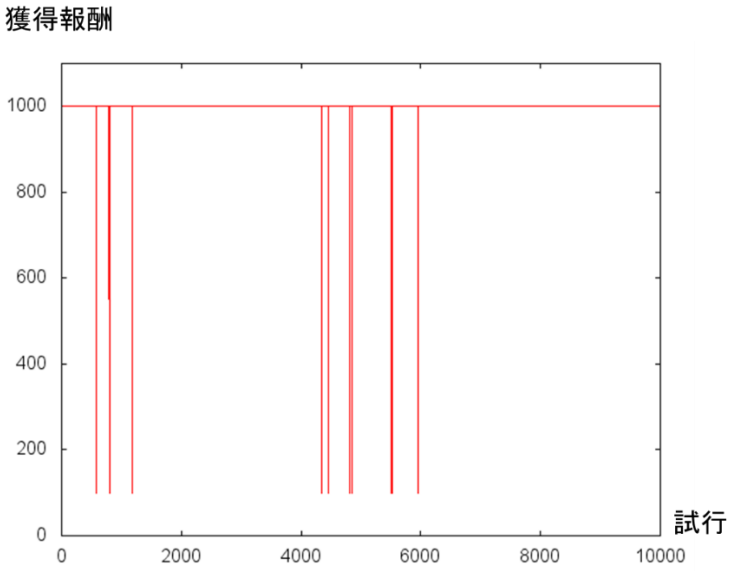

図 4.11 ロボットＢが獲得した報酬(0～10000 試行)

最後 10000 試行で与えられた報酬を図 4.12，図 4.13 に示す.

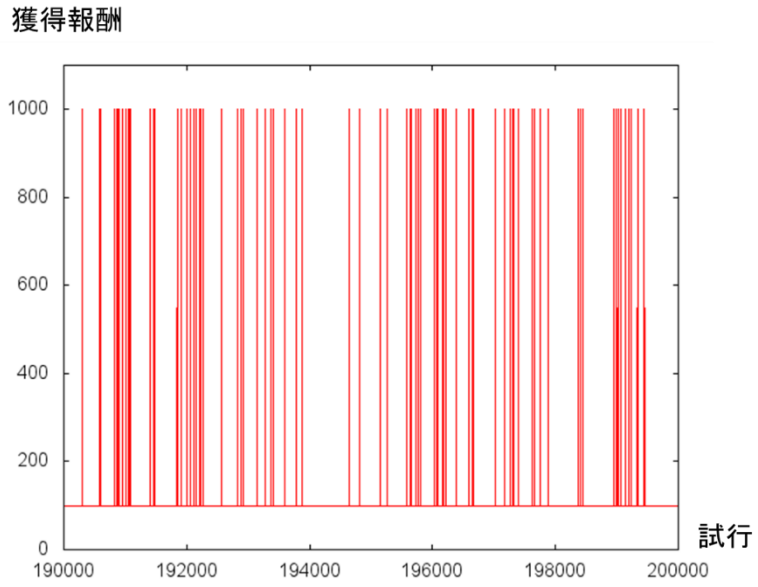

図 4.12 ロボットＡが獲得した報酬(190000～200000 試行)

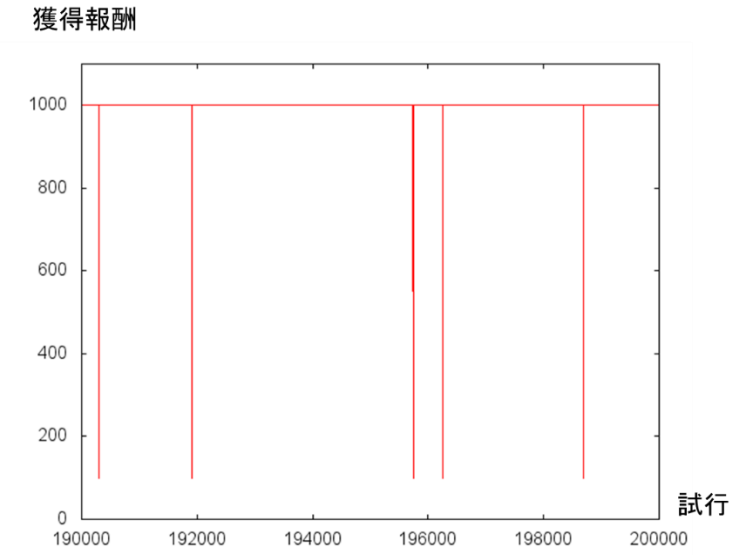

図 4.13 ロボットＢが獲得した報酬(190000～20000 試行)

続いて各ロボットがどれだけサブゴールを通過する行動を学習できるかを調べるため，本実験を 1000 回行って試行毎に与えられた報酬の平均をとった．得られた結果を図 4.14, から図 4.15 に示す.

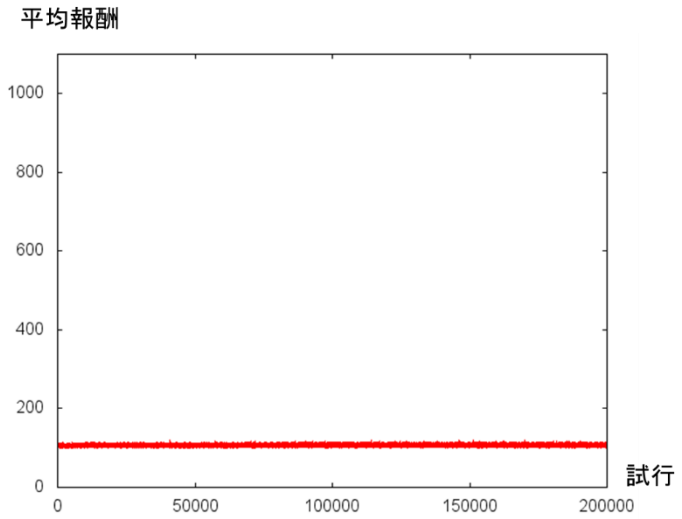

図 4.14 ロボット A が試行毎に獲得した報酬の平均

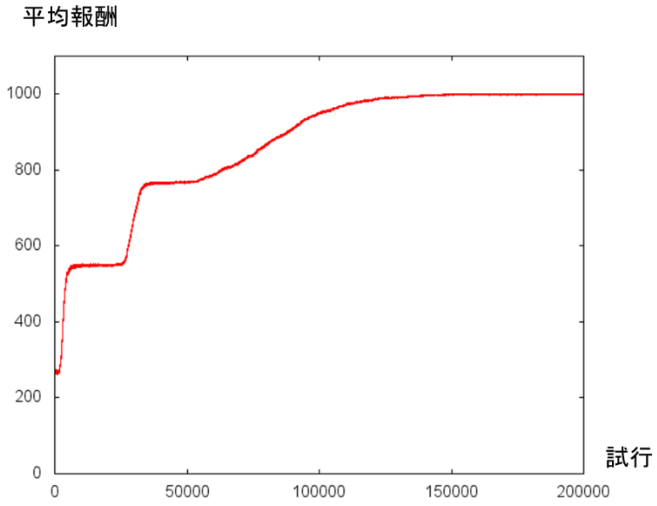

図 4.15 ロボットBが試行毎に獲得した報酬の平均

4.5 考察

4.5.1 ゴールで与えられた報酬についての考察

図 4.4，図 4.6，図 4.8 を見ると，ロボット A は約 8 万試行まではほとんどの試行で 100 の報酬を与えられており，それ以降は与えられる報酬が安定していないように見える． また図 4.10 を見ると，学習開始初期は安定して 100 の報酬を与えられていると考えられ，図 4.12 からは，約 8 万試行以降も与えられる報酬は 100 である試行の方が多いと考えられる．このことから，ロボットAはある試行まではサブゴールを通過しない試行がほとんどだが，それ以降はサブゴールを通過する試行が増える行動を学習していると考えられる．ロボットAに関する結果を通して見ると，ロボットAはサブゴールを通過する行動を安定して学習することができないと考えられる．

図 4.5，図 4.7，図 4.9 を見ると，ロボットBに 1000 の報酬を与えられている試行が多く見られる．また図 4.11 と図 4.13 も合わせて見ると，ロボットAのように与えられる報酬が途中で不安定になることはない．ロボットBに関する結果を通して見ると，ロボットBはサブゴールを通過する行動を学習できると考えられる．

このような結果となった理由として，まずロボット A がサブゴールを通過する行動を学習できないのは，行動選択の方法にあると考えられる．強化学習で用いられる行動選択手法としてはε-greedy 法や softmax 法などがあるが，これらはある状態において選択する行動を行動価値によって決定しており，行動価値の値が最大の行動を選択している．しかし本実験環境のように，サブゴールを通過するためには同じ状態を二度経験しなければならない場合，一つの状態において行動価値が最大の行動が二つ存在してしまう．この場合，図 4.8 のように学習すべき行動が一つに決定しないと考えられる．この

考えを前提に考察すると，ロボット A がサブゴールを通過する行動を学習できていないのは，サブゴールへ到達する前の位置で行動が，サブゴール到達前はサブゴールへ向かう行動となるが，到達し通過した後はゴールへ向かう行動になってしまうためだと考えられる．対してロボットＢはサブゴールを通過後は状態が追加されるため，サブゴール到達前後では同じ位置でも最適な行動が異なるので，サブゴールを通過する行動を学習可能であると考えられる．

4.5.2 平均報酬についての考察

図 4.14 を見ると，ロボット A は与えられた報酬の平均が 100 付近に収束しており，実験を繰り返してもサブゴールを通過する行動は学習できないと言える．また図 4.15 を見ると，二度程平均報酬が急激に増加しており，その後緩やかに上昇を続けている．最終的に 10 万試行を越えたあたりから平均報酬が 1000 に収束している．このことから，ロボットＢは試行を多く繰り返すことで，サブゴールを通過する行動を確実に学習できると言える．

4.5.3 考察のまとめ

4.5.1 節と 4.5.2 節の考察から，提案システムを適用せず，強化学習のみではサブゴールを通過する行動は学習できないが，ロボットＢに適用されている提案システムはサブゴールの発見と通過する行動の学習が可能であると言える．

第 5 章 結論

5.1 全体を通してのまとめ

本研究では，ゴールで与えられる報酬の大きさの差から，ロボットが自律的にサブゴールの発見を行い，発見したサブゴールをクリアしてタスクを達成する行動を学習する学習システムの実現を目標とした．そこで本研究では，経験することによってゴールで得られる報酬が大きくなる状態をサブゴールと定義した．さらにロボットが過去に経験した状態と与えられた報酬の大きさの関係からどの状態がサブゴールであるかを発見し，発見したサブゴールをクリアしているか否かという状態を追加した．これによってサブゴールをクリアする行動を学習し，サブゴールの発見およびクリアする行動の学習を可能とする学習システムを提案した．

また，提案システムによってサブゴールを発見し，発見したサブゴールをクリアしてタスクを達成する行動を学習できることを検証するため，シミュレーション実験を行った．実験結果としては，提案システムを適用したロボットがサブゴールを発見し，さらに状態を追加することによって，発見したサブゴールをクリアしてタスクを達成する行動を学習することが可能であることを示した．このことから，本研究ではゴールで与えられる報酬の差からサブゴールを発見し，発見したサブゴールをクリアする行動を学習するロボットの実現することができたと言える．

5.2 今後の課題

5.2.1 サブゴールの発見においての課題

本論文で提案するシステムでは，サブゴールの探索とサブゴールを経験する行動の学習は別々に行っている．しかし探索と学習を分けるということは，その分学習完了まで時間がかかってしまう．さらに，学習を行う環境が今回行ったシミュレーション実験のように小さいとも限らず，環境が大きくなればなるほど探索にも学習にも時間がかかってしまう．よってサブゴールの探索と行動学習を同時進行し，学習を進めつつもサブゴールを発見できるようなシステムを実現する必要がある．

また本研究ではサブゴールの数は一つで変化せず，一つの状態で固定であると想定したため，サブゴールが複数存在する場合や，数の増減が起こった場合に，サブゴールをクリアする行動を学習できるか否かの検証が済んでいない．

以上をまとめると，サブゴールの探索と学習の同時進行と，サブゴールの数や変化に関わらずに探索を行うことのできるシステムの実現が必要である．

5.2.2 サブゴールの学習においての課題

5.2.1 節で述べたようにサブゴールが複数存在する場合，仮にサブゴールの発見が可能であっても，存在するサブゴールをすべてクリアできる行動を学習できるかどうかがわかっていない．

また今回のシミュレーション実験のように，小さな環境でさえサブゴールをクリアする行動を確実に学習できるまでに，およそ 10 万試行とかなり多くの試行を重ねる必要がある．このことから，本研究で提案する学習システムでは，環境の拡大や複雑化に対応できない可能性がある．

5.2.3 実機への適用

本研究では，実験をシミュレーション実験にて行った．そのため実際のロボットに提案システムを適用した場合に正常に機能するかどうかは確認していない．もし実機に提案システムを適用する場合には，ロボットが動作する環境に応じて蓄積する状態に配慮する必要がある．最終的には実機に提案システムを適用し，提案システムの有用性や適用した場合の問題点などを検証し，実用化を目指す必要がある．

謝辞

本論分を結ぶにあたり，日ごろより懇切なるご指導を賜りました倉重健太郎先生に深く感謝の意を表します．また，ご指導，ご助言をいただきました佐賀聡人先生，畑中雅彦先生，本田泰先生に感謝の意を表します．そして，論文の査読や助言をしていただいた杉本大志さん，渋谷和さん，高泉昇太郎さん，三浦丈典さん，二階堂芳さん，片山和宣君，千葉秀平君に感謝いたします．

参考文献

[1] 辻　敏夫，加藤　壮志，金子　真“人間－ロボット系の追従制御特性”，日本ロボット学会誌，Vol.18，No.2，pp.285-291（2000）

[2] 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）

[3] 齋藤　史倫，福田　敏男“強化学習による実ロボットの運動制御”，日本ロボット学会誌，Vol.13，No.1，pp.82-88（1995）

[4] 小池　康晴，鮫島　和行“強化学習の基礎”
http://www.jnns.org/niss/2000/text/koike2.pdf

[5] 甲斐　孝史，石川　眞澄“強化学習を用いた変動環境下の最短経路探索”，電子情報通信学会信学技報，Vol.109，No.461，pp.119-124(2010)

[6] 前田　陽一郎，花香　敏“Shaping 強化学習を用いた自律エージェントの行動獲得支援手法”，日本知能ファジィ学会誌，Vol.21，No.5，pp.722-733（2009）

[7] 山城　啓秀，上野　敦志，武田　英明“遅れ報酬に基づく遺伝的アルゴリズムによる部分観測マルコフ決定問題の解決手法”，電子情報通信学会論文誌D，Vol.J84-D1，No.9，pp.1635-1647（2001）

[8] 高橋　泰岳，浅田　稔“複数の学習器の階層的構築による行動獲得”，日本ロボット学会誌，Vol.18，No.7，pp.1040-1046(2000)

[9] 森実　克，山田　誠二，豊田　順一“移動ロボットによるサブゴール間巡回行動の学習”，日本ロボット学会誌，Vol.15，No.5，pp.807-810（1997）

[10] 田中　文英，山村　雅幸“MDP 集団の上におけるマルチタスク強化学習”，
電気学会論文誌C（電子・情報・システム部門誌），Vol.123, No.5, pp.1004-1011（2003）