博士論文

同一エージェント間における転移学習を用いた強化学習の高速化に関する研究

A Study on Acceleration of Reinforcement Learning by Transfer Learning for Equivalent Agents

平成24年度
三重大学大学院　工学研究科
博士後期課程　システム工学専攻

高野　敏明

目 次

第1章　序論

1.1　本研究の目的

近年，ロボットは産業用だけでなく，医療用や家庭用として開発されるようになっている [1–3]．これは，ロボットが工場などの整備された環境だけでなく，時々刻々と変化する環境下においても与えられた仕事（タスク）をこなせるようになってきていることを表しており，ロボットの活躍する範囲が広がってきている．将来的には，放射能汚染地域など危険が伴う環境下での作業や介護などの重労働を行うことが期待されている．これらのタスクには，予測困難な事態が生じる可能性が十分にある．このような環境では，柔軟な行動をとることがロボットに要求される．柔軟な行動をロボットがとれるようにプログラムすることは，人間にとって多大な負担が伴う上，予測困難な事態を想定する必要がある．そこで，ロボット自身に直面している事態に合わせた行動を生成させることで，この問題について数多く研究されている [30–33]．また，ロボットに限らず，システムとしても高度で知的な動作が要求されるようになっており，スマートフォンなどに利用されている音声認識技術などは予測困難な事態が生じる可能性があり，インテリジェントなシステムへの応用でも期待される．

ロボットやシステム自身が与えられた環境下で学習を行うことにより，自律的に行動を獲得する手法として，強化学習 [4] が研究されている．強化学習は，工場のような整備された環境だけでなく，住居などの人間の生活空間や宇宙空間などの予測困難な環境下での応用が期待されている．しかし，強化学習は与えられたタスクに対して適切な行動規則を獲得するまでの学習回数が多いという問題が指摘されている [44]．また，タスクに置ける行動目標が同一でも，タスクを行う環境が異なる場合，学習した行動規則により，タスクが達成できないことがある．このような場合，再びタスクを学習する必要があり，この学習にも多くの学習回数が必要となる．以上の理由により，現段階においての実用化が困難である [45]．

強化学習を実用的なものにするため，学習回数を削減する研究が多くなされて

いる．学習回数を削減する研究の例として，学習パラメータの最適化 [22–29]，モデルベース強化学習 [38]，転移学習 [79,80] などが挙げられる．学習パラメータの最適化やモデルベース強化学習は，単位学習回数あたりの学習達成度を高くすることで，学習回数の削減を図る手法である．一方，転移学習は，学習開始前や学習途中から学習達成度をある程度引き上げることで，学習回数の削減を図る手法である．

本研究では，この転移学習に着目し，強化学習における転移学習の手法について議論を行う．

1.2 強化学習

1.2.1 強化学習の概要

強化学習は，状態，行動，そして報酬に関してエージェントとその環境との間の相互作用を定義している形式的な枠組みである [4]．ここで，強化学習において，一般にエージェントとは学習と意思決定を行うものをさす．また，エージェントが相互作用を行う対象を環境と呼ぶ．強化学習においてエージェントは環境と以下のやり取りを行う．

1. エージェントは，環境から状態 s_n を入力として観測する．
2. エージェントは，入力された状態から行動価値関数に基づき，行動 a_n を決定・実行する．
3. エージェントは，状態 s_{n+1} に遷移し，報酬 r_n を環境から受け取る．
4. エージェントは，得られた報酬 r_n をもとに，行動価値関数を更新する．
5. 1〜4 をタスクを達成する（できるようになる）まで繰り返す．

エージェントは環境との相互作用の中で，報酬 r_n を手がかりに，与えられたタスクのすべての状態における行動規則を獲得する．ここで，タスクとは，ある環境下において，あらかじめ定められた所定の状態（以下，ゴールとよぶ）に到達することである．また，行動規則とは，ある状態においてどの行動をすればよいかを表した規則である．一般に，エージェントが与えられたタスクにおいてゴールに到達した際には正の報酬，ある状態において誤った行動をとった際には負の報

酬が与えられる．強化学習のアルゴリズムは，行動価値関数を調整することにより，一行動あたりに受け取る報酬の期待値が大きくなるように定式化されている．

強化学習の環境モデルは，しばしばマルコフ決定過程（Markov Decision Process, MDP）によって定式化される．マルコフ決定過程とは，マルコフ性を持った確率過程を基にし，状態を離散状態として扱う逐次決定過程のことである．また，(単純）マルコフ性とは，次状態 s_{n+1} への状態遷移が現状態 s_n と行動 a_n にのみ依存し，それ以前の状態や行動には関係しない性質をさす．MDP にはいくつかのモデルがあるが [15–17]，ここでは代表的な MDP として，単純 MDP と部分観測 MDP（POMDP）について簡単に説明する．

単純 MDP

単純 MDP は，上で述べた MDP そのものである．単純 MDP は，エージェントが状態遷移に必要な情報をすべて観測できることが条件となる．単純 MDP の例として，コイントスが挙げられる．トスしたコインの表または裏がでる確率はトス直前のコインの状態（表が上か裏が上かなど）と行動（トスの高さや回転のかけ方など）によって決まる，といったものがある．

POMDP

POMDP は，エージェントが状態遷移に必要な情報の一部が観測できる過程のことをさす．POMDP の例として，ポーカーが挙げられる．ポーカーでは，自身手札が観測できる情報であるが，次にどんなカードがくるのか，相手の手札がどのようになっているかを観測することはできない，といったものがある．

本稿では，特に断りをしない限り，この単純 MDP を取り扱うものとする．

1.2.2 アクター・クリティック

強化学習の代表的な手法として，Profit Sharing，Q 学習，アクター・クリティックなどが挙げられる [4–14]．本研究では，強化学習の手法としてアクター・クリティックを取り扱うため，これについて説明する．

アクター・クリティックは，Witten によって提案された強化学習の一種である [6]．Q 学習と比べるとマイナーな手法であるが，近年，Vijay らによってアクター・クリティックの収束性が証明 [10] されてからは次第に注目されるようになった．筆

者はこの経緯を受け，強化学習の手法としてアクター・クリティックを本稿では扱う．アクタークリティックの特徴は，行動を陽に表現しているため，連続値行動に対しても有効である．また，エージェントはアクター (行動器) とクリティック（評価器）により構成されており (図 1.1)，状態価値関数と行動価値関数を区別して保持している (Q 学習では，行動価値関数のみが保持される)．アクターは，環境から状態を観測し，その状態に合わせて行動を選択する．アクターでは，行動価値関数 $p(s_n, a_n)$ を保持しており，この関数は行動優先度とよばれている．クリティックはこの行動の結果として得られる報酬から，TD 誤差 δ(Temporal Difference) を算出し，この TD 誤差により，アクターの選択した行動を評価 · 更新する．クリティックでは，状態価値関数 $V(s)_n$ が保持されている．アクター・クリティックの学習において，エージェントがとりうるすべての状態集合を S，選択可能な行動集合を A とし，状態 $s_n \in S$ における行動 $a_n \in A$ の価値を行動優先度 $p(s_n, a_n)$ という数値で表し，ある状態の推定の価値を状態価値 $V(s_n)$ という数値で表す．アクター・クリティックでは，これらのパラメータの修正を繰り返すことで学習する．行動優先度の修正は，ある状態 s_n において適切な行動優先度 $p(s_n, a_n)$ が最大になるように更新され，状態価値の修正は，ゴールに到達するために通過しなければいけない状態でゴールに近い状態ほど大きくなるように更新される．アクター・クリティックにおける更新式を以下に示す．

$$\delta_n = r + \gamma V(s_{n+1}) - V(s_n) \ , \tag{1.1}$$

$$V(s_n) \leftarrow V(s_n) + \alpha\delta_n \ , \tag{1.2}$$

$$p(s_n, a_n) \leftarrow p(s_n, a_n) + \beta\delta_n \ . \tag{1.3}$$

ここで，式 1.2，式 1.3 において左矢印 ($\leftarrow$) は左辺の変数に右辺の値を代入する操作を表している．また，δ_n は TD 誤差を表し，r_n は状態 s_n において行動 a_n を行った結果，得られる報酬を表す．$V(s_n)$ は状態 s_n における状態価値，s_{n+1} は状態 s_n において行動を a_n を行った結果，遷移した先の状態，$p(s_n, a_n)$ は状態 s_n において，行動 a_n をどれほど優先して行うべきかを表している．γ， α， β はそれぞれ，割引率，学習率，ステップサイズパラメータとよばれるあらかじめ与えられたパラメータ (以降，学習パラメータとよぶ) である ($0 \leq \gamma \leq 1$，$0 \leq \alpha \leq 1$，$0 < \beta$)．割引率 γ は将来獲得予定の報酬を現時点でどれだけ重要視するかの割合を表したものである [83]．また，学習率 α は現時点での $V(s_n)$ と，報酬や遷移した先の状

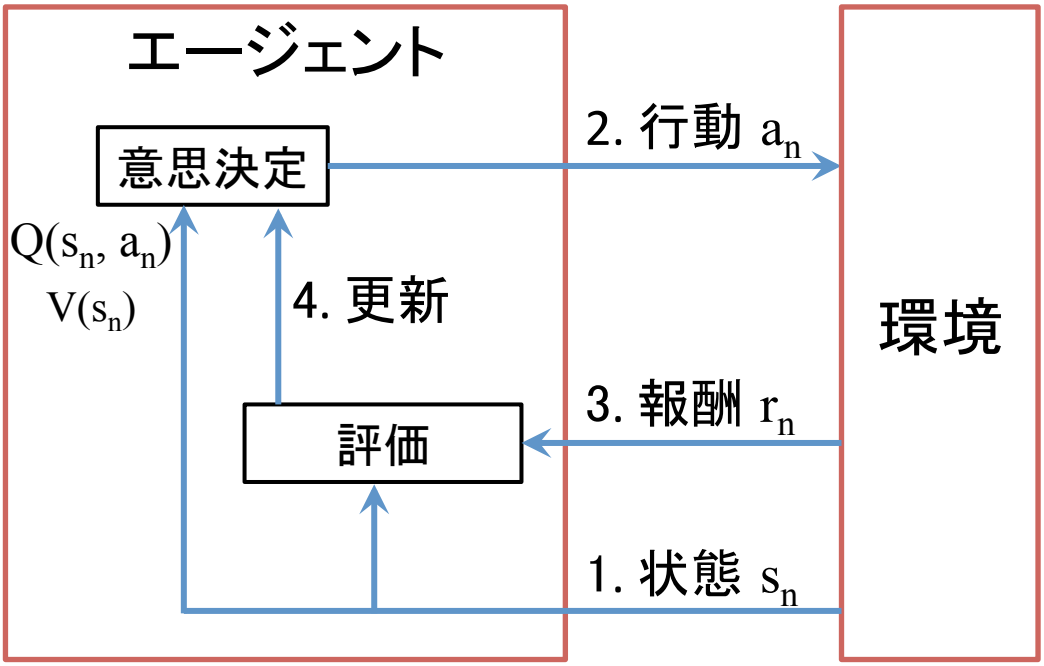


図 1.1: アクター・クリティックの枠組み

態から得られる結果とのバランスを表したものである．同様に，ステップサイズパラメータ β は $p(s_n, a_n)$ において報酬や遷移した先から得られる結果とのバランスを表したものである．アクター・クリティックでは，α，β を十分小さい値にとることで，これらの値にかかわらず 1 つの解に収束する [4] [10].

これらの更新式により，行動優先度と状態価値の更新を繰り返すことで，各状態における適切な行動の行動優先度が最大になるように修正する．ここで，適切な行動とは，アクター・クリティックではゴールに向かう行動のことを指す．最終的に，各状態における最大の行動優先度を持つ行動が，与えられたタスクにおける各状態での行動規則となる．

問題点としては，強化学習は試行錯誤により，適切な行動を学習するため，学習が完了するまでには多数回の学習が必要になることが広く知られている [45].

1.2.3 行動選択手法

強化学習では，学習パラメータを適切に設定し，十分な学習を行うことでタスクを達成する知識を獲得することができる．その際，エージェントの行動決定の方法によって学習の進み方にも違い現れる．代表的な行動選択手法を以下に示す．

- greedy 法
- $\epsilon-$greedy 法
- ソフトマックス手法

各行動選択手法について説明する．

greedy 法とは，現状態 s_n において最大の行動価値を持つ行動 a_n を決定的に選択する手法である．強化学習では，各状態における適切な行動価値を最大にするように学習するため，greedy 法は最も単純な行動選択手法である．

$\epsilon-$greedy 法は，確率 $1-\epsilon$ で greedy 法によって行動を選択肢，確率 ϵ で行動価値関数によらずランダムに行動を選択する手法である．この手法は確率 ϵ で greedy 法とは異なった行動を選択できるため，エージェントは greedy 法よりも探索的になる．また，$\epsilon = 0$ のとき，$\epsilon-$greedy 法は，greedy 法と等価となる．

ソフトマックス手法はルールの価値の比によって確率的に行動を選択する手法である．比の計算としては，ボルツマン選択やルーレット選択などが用いられることが多い．ボルツマン選択は，状態 s_n における行動 a_n を選択する確率 $p(a_n|s_n)$ は以下の式によって求める．

$$p(a_n|s_n) = \frac{\exp(p(s_n, a_n)/T)}{\Sigma_b \exp(p(s_n, b)/T)} \tag{1.4}$$

この式より，大きな行動価値を持つ行動ほど選択される確率が高く，小さな行動価値を持つ行動ほど選択される確率は低くなる．なお，T は温度パラメータとよばれ，$T \to 0$ のとき，greedy 法と等価になる [4]．

いずれの行動選択手法においても，学習の進行を促すようなパラメータの設定は難しい．また，現状ではどの行動手法をとるべきかという明確な指針は現状では示されていない [83]．

本論文では，検証実験にて $\epsilon-$greedy 法とボルツマン選択を使用している．

1.3 強化学習における学習の高速化

強化学習により，タスクに対する行動規則を学習によって自律的に獲得できる．その一方，学習回数が多いという問題点が指摘されており，現段階では，実用化が困難であるといわれている [45]．実用化に向け，この問題点を解消するための

研究が多くなされている [22–43]．本節では，学習の高速化に関する従来研究を紹介する．

学習の高速化に関する従来研究のねらいは次の 2 つに分類できる．

1. 現在行っているタスクにおける単位学習回数あたりの学習達成度の上昇率を上げる [22–38]．

2. 現在行っているタスクの学習開始時や学習途中で学習達成度を引き上げる [39–43, 46–78]．

ここで，学習達成度とは，適切な行動をとることができる状態の全状態に対する割合のことをさす．学習達成度が 0% に近い程適切な行動をとることができる状態が少なく，学習達成度が 100% に近いほど，適切な行動をとることができる状態が多いことを意味する．前者は，学習の仕方を工夫することで，学習達成度の上昇率を上げるものである．一方，後者は学習とは別の方法で学習達成度を引き上げるものである．それぞれのねらいのもとでの従来研究を簡単に説明する．

まず，単位学習回数あたりの学習達成度の上昇をねらいとする従来研究について紹介する．これをねらいとする従来研究は数多くある [22–38]．ここでは，パラメータ設定の最適化 [22–29] とモデルベース強化学習 [38] の 2 つを紹介する．

- パラメータの最適化
  パラメータ設定の最適化は，強化学習の学習パラメータである学習率 α や割引率 γ などを遺伝的アルゴリズムによって，最適な数値となるように設定・調整する手法である．これは，学習パラメータを適切に設定することで，学習によって変化するパラメータが適切に更新され，学習回数を削減するねらいがある．

- モデルベース強化学習
  モデルベース強化学習は，エージェントの内部に環境モデルを作り，実際の学習とモデル内の学習の 2 種類の学習を行うことで，学習によって変化するパラメータを更新する手法である．これはいわば，人がイメージトレーニングをした上で動作を行うのと同様に，実際の行動は 1 回でも，複数回のパラメータ更新を行うことができる．これにより学習回数の削減をねらうものである．また，学習 1 回は，エージェントが状態を観測する，行動を選択・実行する，その結果として報酬を受け取り，各パラメータを更新する，までを意味する．

次に後者をねらいとする従来研究として転移学習を紹介する．転移学習は，過去に獲得した知識を，これから学習するタスクに利用する手法である．これは，現在行っているタスクの学習開始時や学習途中から過去に経験した知識を利用することで，適切な行動をとることができる状態を多くし，学習回数を削減するねらいがある．転移学習の研究も多く研究されている [39–43,46–78]．ここでは，PRQ-Learning (Policy Reuse in Q-Leanring) [71] と概念学習 [97] の 2 つを紹介する．

- PRQ-Learning
  PRQ-Leerning では，実際に過去の知識を用いて現在のタスクで行動をさせ，その結果から過去の知識として最も適した過去の知識を選ぶ．それと同時に，現在のタスクの学習を行う手法を提案している．このとき，エージェントの行動は，確率 φ で過去の知識に基づいた行動を行い，確率 $1-\varphi$ で学習中の知識を用いて $\epsilon-$greedy 法に基づいて行動を行うというものである．

- 強化学習結果の再構築への概念学習の適用
  概念学習を導入した事例では，学習開始前に概念学習を用いて過去のタスクと現在のタスクを比べ，過去の知識が利用できるかを判断する手法を提案している．この手法では，学習開始前の時点でタスクの概念が認知できるという前提で，過去のタスクと現在のタスクとの概念を比較して，類似しているタスクの知識を転移するというものである．

われわれ人間は，現在直面している問題を学習するときには，0 から学習するのではなく，過去に学習したタスクで獲得した知識を利用している．そこで，強化学習においても，ある環境で現在行おうとしているタスクにおける学習回数を削減するためには，0 から学習を開始するのではなく，過去のタスクで獲得した知識を利用することが自然である．そこで，[71] と [97] を，人間の転移学習を機械学習へ応用した例として取り上げた．前者は転移する知識の使い方についての論文，後者は，転移する知識を選ぶ方法についての論文である．[39–43, 46–78] と多くの転移学習を用いた機械学習の手法が提案されているが，決定的な手法は未だ提案されておらず，議論すべき点が残されていると筆者は考えている．そのため，本研究では，この転移学習に着目し，強化学習の高速化を図る手法について検討を重ねることとする．次章にて，この転移学習についてもう少し説明をする．

1.4　本論文の構成

本論文は，上記の目的の下に行われた研究をまとめたものであり，全 5 章で構成される．

第 1 章では，本研究で対象とした強化学習とその問題点について説明した．また，本研究の目的について述べた．

第 2 章では，本論文で着目した転移学習について，転移学習の流れについて簡単に説明したのち，強化学習における転移学習の議論点と周辺研究について簡単にまとめる．

第 3 章では，転移学習を用いた強化学習のモデルとして異遷移同目的ドメインに焦点をあて，このモデルに合わせた転移学習を用いた強化学習のアルゴリズムを提案する．また，同アルゴリズムの有効性について簡単な実験を通して検証を行う．

第 4 章では，転移学習を用いた強化学習のモデルとして同遷移異目的ドメインに焦点をあて，このモデルに合わせた転移学習を用いた強化学習のアルゴリズムを提案する．また，同アルゴリズムの有効性についても実験を通して検証を行う．

第 5 章で，本研究によって得られた成果をまとめ，今後の課題について触れる．

なお，参考文献は，最後にまとめて掲載する．

第2章　強化学習における転移学習

強化学習における転移学習の利用に関する研究は他にも議論がなされている [39–43, 46–78]．本章では，転移学習の概要を説明し，周辺研究について俯瞰する．また，本論文において議論する点について説明する．

2.1　転移学習の歴史と背景

転移学習（Transfer Learning）という語は，聞き慣れないかも知れないが，それほど新しいものでもない．転移という語は，もともと教育心理学の分野で使われていたものである [89–93]．転移学習の研究は古くから存在し，Thorndike らによって 100 年以上前から研究が始まっている [88]．人間における転移とは，過去の体験・経験や書物から得るあるいは教授された知識などを応用することをいう．つまりは，知識の再利用することを指している．教育心理学の分野においては，生徒が転移を行えるまで指導することが目標であり，教育者にとって大きな課題である．そのため，日々教育者は，生徒に知識の転移ができるよう，指導方法に工夫を重ねている [89–93]．これについて，Haskell はこの転移のレベルを 6 階層に，知識を 5 種類に分けて説明している [94]．転移のレベルは以下に示すとおりに分けている．

1. Nonspecific Transfer
   すべての学習において起こりうることで，過去の学習と現在の学習とを結びつけること．無意識にも行われていることもある．

2. Application Transfer
   特定の状況で学習したことを同じような状況で知識を転移する場合に起こる．

3. Context Transfer
   特定の状況において学習した過去の知識が，条件を少し変えた現在のタスク

がその知識を用いることで達成できるとは限らない．こうした状況で起こる転移のことを指す．

4. Near Transfer
   類似した知識で問題を解決するときに起こる．

5. Far Transfer
   状況の把握や類推をするときに起こる..

6. Displacement or Creative Transfer
   新しいものと古いものとの間で，類似点のインタラクションが起こり，新しい概念を作りだす場合に起こる．

このように，6 つの転移レベルから転移学習は構成されている．概念的な話では，分かりにくいと思うので，具体的な例を挙げる．

- 普通自動車の運転技術で，大型自動車の運転をする．
  ハンドルをきるタイミングやブレーキを踏むタイミングが異なるが，うまく運転することができる．

- 試験対策として，過去の問題集を解いて練習する．
  問題の設定や値が異なっていても，出題傾向をつかみ，高得点をあげることができる．

- 先輩のレポートを参考にして，新しいレポートを作成する．
  レポート課題の内容が若干変化していても，先輩のレポートから利用できる部分と自分で新しく考察した部分とをうまく組合せたレポートを作成する．

このように，人間の生活の中には，転移学習が行われている部分が非常に多くある．前述した例からも分かるように，人間の「応用力」と呼べる多くが知識の転移により，成し得ているものであるといえる．

本研究の対象は，この人間の応用力（転移）の機構を強化学習の分野に取り入れようとする試みである．次節において，強化学習に転移学習を適用することについて説明する．

2.2　強化学習における転移学習の議論点

強化学習は，環境との相互作用により自律的に目的の知識を獲得することができるが，目的の知識を獲得するまでに多くの試行錯誤が必要である．そのため，学習に時間がかかるという問題が指摘されている．そこで，この学習の高速化を図るため，転移学習が着目され始めている．転移学習については，前節でも説明したように，過去に獲得した知識を現在直面している問題に利用することで，学習の高速化を図る手法である．具体的には，知識を転移（応用）すること，転移した後の学習により，現在直面している問題を解決する手法である．

本研究は，転移学習を強化学習に適用することにより，学習の高速化を図っている．図 2.1 で，本研究における強化学習と転移学習の関係について示す．強化学習では，ロボットの動作獲得やパラメータ調節による学習の高速化，仮想学習を用いた学習の高速化などが研究されている．一方，転移学習は，教育心理学の分野で広く知られている．転移学習を機械学習に用いる例として，CBR(Case Based Learning) や ANN(Artificial Neural Network) などの分野で用いられている．なお，以降，転移学習とよぶ時は，特に断りがない限り転移学習を用いた強化学習を意味することとする．

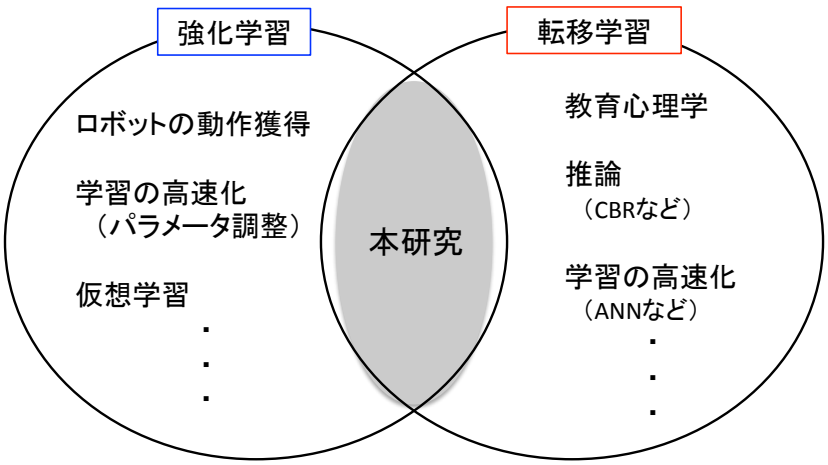


図 2.1: 本研究における研究領域の解説

図 2.2 に転移学習を用いた強化学習のイメージ図を示す．転移学習では，まず事前に複数のタスクをエージェントに学習させ，獲得した知識をデータベースとして構築する．そして，現在のタスクに対して，データベースの中から必要な知識，あるいは，情報をエージェントに転移する．最後に強化学習により，現在のタス

クを達成するための行動規則へと調整を行う．これが強化学習における転移学習の一般的な流れである．このとき，複数の知識を転移させる知識として選択してもかまわない．また，部分的な知識の転移も可能である．これについては，別途議論をする必要があるのだが，本論文では，1つの知識を転移させるとして話を進める．

強化学習における転移学習として重要なことは以下の3点である．

1. 知識データベースの構築

2. 転移させる知識の選択

3. 知識の転移方法

1については，まず転移するために必要な過去の経験・知識をデータベースに蓄積することである．利用できる知識は多ければ多いほど転移学習による学習の高速化が期待できる．そのため，多くの知識をデータベースに蓄積する必要がある．しかし，メモリ容量などの制限があるため，データベースに蓄積する知識は取捨選択し，ある程度一般化された知識が蓄積されるのが好ましい．ある程度一般化された知識がデータベースに蓄積されるよう，議論を行う必要がある．

2については，1で構築したデータベースから現在のタスクを効率的に進めることのできる知識を選択することである．良い知識が選択された場合，エージェントはタスクを達成するための知識を学習することなく獲得できるため，学習が速く完了する．しかし，悪い知識を選択された場合，エージェントは悪い知識を修正し，なおかつタスクを達成するための学習が必要となる．そのため，悪い知識が選択された場合，転移学習により学習が遅くなる．エージェントが良い知識を選択し，学習の高速化につながる知識を選択できるよう，議論する必要がある．

3については，2で選ばれた知識の中で，さらに細かなレベルで転移させる情報を吟味することである．知識を構成する情報には，(a) タスク達成に関係する正しい情報，(b) タスク達成には無関係の情報，(c) タスク達成に関係する間違った情報，の3種類が考えら得る．2において，(a) で構成される知識が選択されるのが理想である．しかし，そのような場合はまれで，多くの場合，(a)，(b)，(c) が含まれる知識である．そのため，選ばれた知識の中で (b)，(c)（少なくとも (c) を）を取り除いた転移学習が行われるよう議論する必要がある．

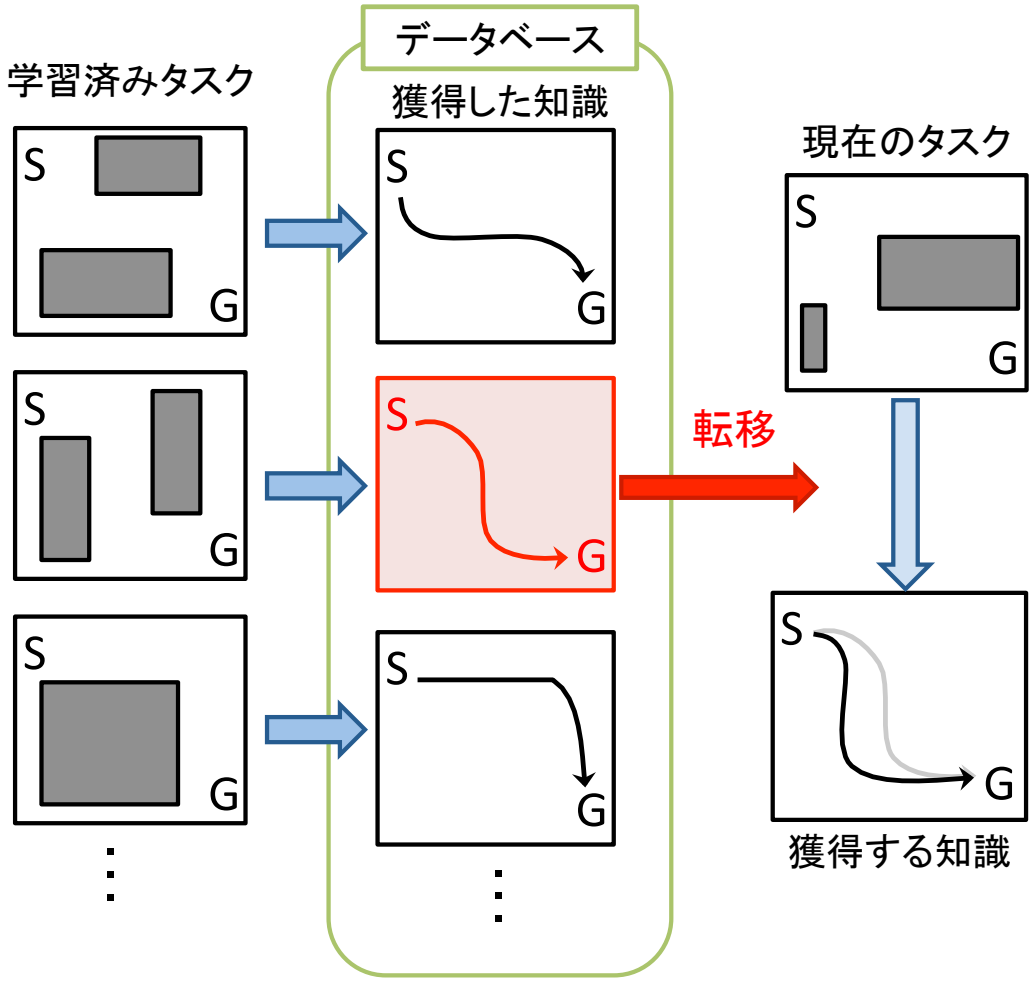


図 2.2: 転移学習を用いた強化学習のイメージ図

2.3 周辺研究の俯瞰

転移学習では，すべての知識が効果的に転移できるわけではない．例えば，迷路を走破するタスクとボールを投げるタスクではタスクとして異なるため，これらのタスクにおいてエージェント間で転移学習を行ったところで有効な知識が得られるとは限らない．これはエージェントについても同様である．用途の異なるエージェントを準備して，それぞれのエージェントが保持している知識を別のエージェントに利用することは一般的にできない．しかし，タスクやエージェントの関係について条件を設定することでエージェント間でにおいて転移学習を行うことが可能となる．このときのタスクとエージェントの情報はドメインとよばれている [82]．本節では，このドメインについて説明し，周辺研究について俯瞰する．

2.3.1 ドメイン

この項では，転移学習を行う際の前提条件となるドメインについて説明する．ドメインは，エージェントに挑戦させる問題をモデルとして表すものである．このドメインは，エージェントと環境により構成され，モデルの構成要素は，エージェントが観測できるすべての状態，エージェントがとりうるすべての行動，状態遷移確率，環境から与えられる報酬など複数の要素により構成されている．

強化学習は基本的に，マルコフ決定過程を前提としており，このマルコフ決定過程のドメインは一般的に次のように表される．

マルコフ決定過程 マルコフ決定過程は，$< S$，A，T，$R >$の組として表すことができる．ここで，S はエージェントが観測できるすべての状態，A はエージェントが選択できるすべての行動，T はある状態 s_n において行動 a_n を行った際，状態 s_{n+1} に遷移する確率，R は状態遷移によってエージェントが受け取る報酬を意味する．

このように問題が持つ要素を合わせてドメインとよび，強化学習におけるマルコフ決定過程においては，$D =< S$，A，T，$R >$ と表される．

ここで，強化学習において転移学習を適用する際の関係について考える．転移学習では，過去のタスクと現在のタスクが存在するので，それぞれを $< S_{\mathrm{s}}$，A_{s}，T_{s}，$R_{\mathrm{s}} >$，$< S_{\mathrm{t}}$，A_{t}，T_{t}，$R_{\mathrm{t}} >$ と区別する．添字 s は転移元を表し，添字 t は転移先を表している．転移学習では，これらの関係（例えば，$S_{\mathrm{s}} = S_{\mathrm{t}}$ 又は，$S_{\mathrm{s}} \rightarrow S_{\mathrm{t}}$ な

ど）についてエージェントが把握していなければ，別の状態や行動へ知識を転移させてしまう可能性がある．

また，これら S，A，T，R などのドメインの組合せの中で，どれか 1 つでも異なる要素がある場合，異なるモデルとして，転移学習を考える必要がある．なぜならば，$S_{\mathrm{s}} \neq S_{\mathrm{t}}$ の場合，エージェントが観測する状態数や状態そのものが違うことになる．$A_{\mathrm{s}} \neq A_{\mathrm{t}}$ の場合は，エージェントが選択できる行動が違うことになる．どれか 1 つでも要素が異なる場合は，予期せぬ転移が発生する恐れがある．そのため，多くの研究者は過去のタスクと現在のタスクのドメイン間においていくつかの定義を行い，転移学習を用いた強化学習について議論を重ねてきた．次項では，従来の研究についてもう少し詳しく述べる．

2.3.2 ドメインラベルによる転移学習の違い

強化学習における転移学習では，現在のタスクに有効な知識を転移が早いほど学習の高速化の効果が大きい．学習開始前あるいは学習開始直後に転移させる知識がわかるならば，その時点から転移学習を行うことが望ましい．データベースから知識を転移する際，受け手となる現在の直面している問題に対するドメインをターゲットドメイン，送り手となるデータベースに存在する知識に対するドメインをソースドメインとよばれている．これらのドメインについて神嶌は「訓練事例に教示情報（出力情報）があるかどうかによって，4 種類の設定が考えられる」と述べている [82]．

(1) 帰納転移学習

ソースドメインにもターゲットドメインにもラベルが付された状況下における転移学習を指す．帰納転移学習の研究はもっとも多い．双方のラベルが既知のため，現在の直面しているタスクに対していかに最適な知識の使い方をするのかが問題となる．

(2) トランスダクティブ転移学習ソースドメインにはラベルが付されているがターゲットドメインにはラベルが付されていない状況下における転移学習を指す．これは，(1) の次に研究が多い．トランスダクティブ転移学習は，いかにしてラベルのない目標ドメインのデ―タの適切なラベルを予測するかが問題である．

(3) 自己教示学習ソースドメインにはラベルが付されていないが，ターゲットドメインにはラベルが付されている状況下における転移学習を指す．自己教示学習については，いかに学習が終了したデータに対してラベルを付けるのかが問題となっている．

(4) 教師なし転移学習ソースドメインにもターゲットドメインにもラベルが付されていない状況下における転移学習を指す．教師なし転移学習では，どちらにもラベルがないため，いかにして双方のラベルを予測するかが問題となっている．

これは，一般的な転移学習において行われている研究であるが，強化学習における転移学習においてもこれらの議論が必要となることはいうまでもない．

本論文では，ソースドメインにはラベルが付されているがターゲットドメインにはラベルが付されていないトランスダクティブ転移学習に焦点をあてて議論を行う．

2.3.3 エージェントによる転移学習の違い

この項では，エージェントの違いに着目し，強化学習おける転移学習の周辺研究がどのようになっているかを簡単に説明する．

強化学習における転移学習では，どんなエージェントに知識を転移するのかが問題となる．強化学習のエージェントは認知する状態 S と行動 A がエージェント固有である．そのため，以下に示すように，過去に利用したエージェントと現在利用しているエージェントとの関係により転移学習に違いが生じる．

◆ 同一エージェント間での転移学習 [46–56]

エージェントが認知できる状態，選択できる行動が過去に利用したエージェントと現在利用しているエージェント同等の機能を有するエージェント間での転移を指している．ここで,「同等の機能」とは，同じエージェントではなくても，認知できる状態，行動が同じという意味である．

同一エージェント間での転移学習においては，前述した 3 点について議論する必要がある．従来の研究では，転移させる知識が人間によって選ばれていることを前提とした議論がなされている．

◆ 類似エージェント間での転移学習 [57–70]

エージェントが認知できる状態，選択できる行動が過去に利用したエージェントと現在利用しているエージェントとは異なるエージェント間での転移を指している．ここでは，状態あるいは行動のどちらかが異なる場合も含まれる．また,「類似」とはタスクの性質として同じという意味を指している．

類似エージェント間での転移学習は，3 点の同一エージェント間における転移学習の議論に加え，エージェント間の写像関係を把握する手法についての議論する必要となる．これは，エージェント間で状態数あるいは行動数が異なるので，対応する状態や行動に転移できない場合が発生するためである．従来の研究では，転移させる知識は人間により既に選ばれており，エージェント間における状態・行動の写像関係がすでに与えられている [57–63] と自律的にエージェント間における状態・行動の写像関係を類推する場合 [64–70] との 2 つの状況について検討が行われている．

◆ マルチエージェントへの適用 [83–87]

マルチエージェントは，複数の同一エージェントあるいは類似エージェントにより一つのタスクを達成するものである．ここでは，どのエージェントも現在利用しているエージェントだが，それぞれのエージェントが獲得した情報を共有することは転移の一種ととらえられる．

マルチエージェントでは，同一エージェントと類似エージェント間での転移学習に加え，学習中の知識が互いに共有されることを考慮する必要がある．

同一エージェント間における転移学習と類似エージェント間での転移学習では，過去の知識を現在の学習に利用する転移学習であるが，マルチエージェントにおける転移学習では，過去という概念ではなく，それぞれのエージェントが経験したリアルタイムの情報を共有するものである．これは，広く見れば転移学習であり，それぞれのエージェントにより獲得された情報が知識データベースとみなし，転移学習を行っている．それぞれのエージェント間での転移学習にはそれぞれの課題がある．ここで，エージェントの違いによる従来の研究について [79] でまとめられているため，これを参考にエージェントの違いによる転移学習の従来研究について説明する．表 2.1 には同一エージェント間における転移学習，表 2.2 には，類似エージェント間における転移学習についてまとめてある．なお，表 2.1，2.2 内の各文字についての説明を表 2.3 にまとめる．表 2.1，2.2 を見ると知識の選択につい

ては，人によって選択されるかすべての知識を使う手法がほとんどで，エージェントにより転移学習を行うための知識の選択方法については述べられていない．また，エージェントにより転移学習を行うための知識の選択手法として [95]– [100] などが提案されているが，エージェントの違いに着目して，知識の転移を検討しているものはほとんどない．

表 2.1: 同一エージェントによる転移学習の従来研究

文献	タスクの違い	知識の選択	学習法	検証事項
[46]	r, s	h	H	tr
[47]	s_i	h	TD	tt
[48]	r	all	H	tt
[49]	r	all	MB	tr
[50]	r, s	h	Batch	tt
[52]	s, t	all	TD	tt, tr
[53]	s,t	h	TD	tr
[54]	t	h	TD	tt
[55]	s_f, t	h	TD	tr
[56]	r	all	TD	tr

同一エージェント間での転移学習では，知識データベース内の知識はすべて現在のエージェントと観測される状態が同じかつ選択できる行動も同じであるため，知識をそのままの形で転移することができる．ここで問題となるのは，

- どのように状態が遷移するのか，
- なにが目的とされているのか，
- いつ報酬がもらえるのか，

といった点について考える必要がある．これは，転移学習を強化学習に用いる際の基本的な検討事項である．

次に，類似エージェント間における転移学習は，知識データベース内における知識が現在のエージェントがもつ形と異なっている場合である．このときの注意点に，前述した同一エージェント間における転移学習の検討も必要になる．その上，過去のエージェントと現在のエージェントにおいて，

表 2.2: 類似エージェントによる転移学習の従来研究

文献	タスクの違い	知識の選択	学習法	検証事項
[57]	a, v	h	TD	tr
[58]	#	h	RRL	tr
[59]	#	h	LP	tr
[60]	p	h	TD	tr
[62]	#	h	RRL	tt, tr
[63]	s, t	h	TD, CBR	tr
[64]	a, v	h	PS	tt
[70]	a, r, v	h	TD	tr

表 2.3: 表 2.1，2.2 の各記号の意味について

タスクの違い	学習法
a: 行動空間	H: 階層型強化学習
p: 問題空間	TD: TD 学習
r: 報酬関数	MB: モデルベース強化学習
s_i: スタート	Batch: バッチ学習
s_f: ゴール	RRL: 関係強化学習
t: 状態遷移確率	LP: 線形計画法
v: 状態価値	PS: 方策探索法
# オブジェクトの数	
	検証事項
知識の選択	tr: 総報酬
h: 人による選択	tt: 学習時間
all: すべての過去の知識	

- 状態空間の関係,
- 行動空間の関係,

を把握する必要がある．これは，過去のエージェントと現在のエージェントとで少なくともどちらか一方の空間が異なると，知識の転移を行った際に設定されていない状態や行動を転移しようすることが容易に想像できる．また，転移ができたとしても知識とは異なった行動をとるため，転移学習の意味をなさない．そのため，エージェント間の状態空間や行動空間を推定することが必要となる．

最後にマルチエージェントに応用する場合，上記2つの同一エージェント間における転移学習と類似エージェント間における転移学習に加え，リアルタイムでの情報共有である点をふまえる必要がある．これは，学習中のエージェントであれば，不確定の情報を含んでいることを想定した転移を行う必要がある．また，他のエージェントと協調した行動が要求されるため，単純な知識の転移では，協調行動とはならないことが容易に想像できる．

このように，エージェントの違いによる，転移学習を用いた強化学習には多くの議論が必要である．そして，決定的な手法は提案されていない．そこで，本論文では，議論を簡単にした同一エージェント間における転移学習に焦点をあて，転移学習を用いた強化学習の高速化に関する議論を行う．

2.4 本論文における研究対象

本節では，本論文において着目する対象についてまとめる．

転移学習を用いた強化学習には，議論の余地が多く残されていると筆者は考えている．例えば，どういうエージェント間についての議論を行うのか，環境に違いはあるのか，タスクに違いはあるのか，ドメインラベルはどのようになっているか，などさまざまなケースが考えられる．転移学習の関係を大まかに描くと図2.3のように表せる．類似エージェント間とは，センサーなどの観測状態や行動の性質の概念が同じであるが，性能が異なることエージェント間のことを指している．同様に，類似環境は，災害前後の地形など，大部分の環境が同じということを意味している．また，類似タスクは，タスクの概念レベルで同じタスクのことをさしている．転移学習を用いた強化学習には，多くのケースが考えられるため，本論文では，図2.3の中から，同一エージェントに的を絞り，類似環境における同一タスク，同一環境における類似タスクの2点について議論することとする．

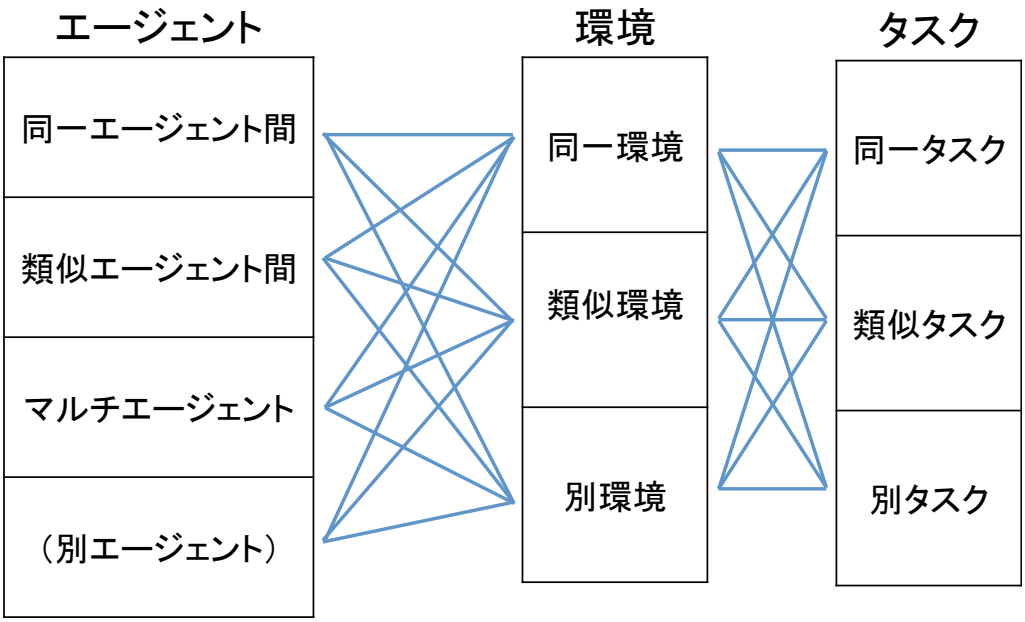


図 2.3: 転移学習を用いた強化学習の関係図

転移学習を強化学習に用いて学習の高速化を検討する際の問題の一つに，学習時間の定義問題がある．これはいつを基準として転移学習により現在学習している環境を学習し終えるまでの時間とするかである．これには 2 つの学習時間の考え方がある．

- 過去の学習時間と現在の学習時間を合わせた学習時間,
- 単純に現在直面しているタスクだけの学習時間,

転移学習の研究の多くは，後者を議論している．また，本論文では，以前に経験した知識が使うことを前提とするため，後者における学習の高速化について議論を行う．

強化学習には多くの手法が存在する．Profit Sharing や Q 学習，アクター・クリティックなどさまざまな手法がある．本論文ではアクター・クリティックに着目する．アクター・クリティックの特徴として，状態行動対の価値観数だけでなく状態価値関数を持つこと，行動方策を陽に示していることなどが挙げられる．これらの特徴は，効果的な転移学習のためには，パラメータが多い方が転移のレベルのバリエーションが多くできると考えられる．そのため，本論文では，強化学習の手法として，アクター・クリティックを使用する．

前節で示したように，一般的な強化学習のドメインはマルコフ決定過程に基づいている．そのため，ドメインは$D < S$，A，T，$R >$となるが，本論文では，エージェントの目標を陽に表現できるよう，この目標をドメインとして新しく加え，ドメインを$D' < S$，A，T，R，$G >$と変更する．この理由については，3章において説明する．

転移学習については前述したように，

◆ 同一エージェント間での転移学習，

◆ 類似エージェント間での転移学習，

◆ マルチエージェントでの適用，

と3種類が考えられる．本論文では，転移学習を用いた強化学習の基礎を検討するため，同一エージェント間での転移学習に焦点をあてて議論を行う．また，本論文における転移学習は学習中に行ない，エージェントは現在のタスクについての事前情報を持たず，過去に獲得した知識は，強化学習による行動価値関数として，データベースに蓄積されているものを議論の対象とする．

2.5 結び

本章では，転移学習を用いた強化学習について触れた．まずは，転移学習の歴史として人の転移について簡単に記述した．そこで，転移のレベルとして6段階のレベルについて説明した．人の転移は複雑でこれとは別に5種類の知識と14種にもおよぶ転移の区分が存在するが，強化学習における転移学習では，そこまで考慮する必要がないため割愛した．これは，エージェントが学習により獲得した知識は，数値データとして記録されるため，知識の種類を定めることが難しいためである．

次に，転移学習を強化学習に利用するにあたり，議論すべき点について簡単にまとめた．エージェントやタスクを表すドメインについて説明を行い，さまざまなドメインにおける転移学習を示し，簡単な周辺研究の俯瞰を行った．また，転移学習を用いた強化学習における学習の高速化の観点において重要な点は，

1. 知識データベースの構築，

2. 転移させる知識の選択，

3. 知識の転移方法,

の 3 点であると考えられる．(1) は，転移する知識がなければ転移学習ができないので，転移するための知識をデータベースとして保有すること．このとき，すべての経験をデータベースに蓄積することは，メモリ容量の不足を招くので現実的ではないので，複数のタスクに利用できかつ学習の高速化につながる情報を多く持った知識をデータベースに蓄積することが望ましい．ただし，本論文ではこれについての議論はしていない．また，(2) で知識を選ぶ際，豊富な知識の中から不適切な知識を選ぶことは障害にもなるので，できる限り少なくする必要がある．(2) は，データベース内のどの知識でも使えばいいというものではなく，現在のタスクの学習に必要な情報を含む知識を選ぶことが，強化学習を高速化するために必要である．(3) は，(2) で選ばれた知識のすべてが現在のタスクの学習に必要な情報であることは珍しい．そのため，選ばれた知識の中でも現在の学習進行を促進するような情報を選び転移することで，学習の高速化につながると考えられる．本稿では，特に (2)，(3) についての議論を行う．

第3章　異遷移同目的ドメインにおける転移学習

本章では，データベースにある知識の中で，現在直面している学習の進行を促進する知識を選択するための手法について検討を行う．検討を行う前の前提として，学習中に現在のタスクに転移すべき知識を選ぶこと，エージェントには現在のタスクに関する事前情報を与えないこと，エージェントは転移学習用に別途特殊なセンサー等で観測をしないことを前提とする．

まず，3.1 では，この章で対象とするドメインについて定義する．3.2 では，一般的な強化学習で得られる行動価値関数からタスクの特徴を抽出する方法について検討する．3.3 では，行動価値関数とは別の情報利用してタスクの特徴を抽出する手法について検討を行う．そして，3.4 では，抽出された特徴により，学習の高速化につながる知識が選ばれているのかを簡単な実験を通して検証を行う．3.5 では，3.3 により選ばれた知識の利用の仕方について検討を行い一連の提案をアルゴリズムとしてまとめる．そして 3.6 で，3.3 と 3.5 により提案された手法を転移学習を用いた強化学習の一連のアルゴリズムとしてまとめる．3.7 では 3.6 のアルゴリズムにより学習の高速化の効果について簡単な実験により検証を行う．そして，最後に 3.8 にて本章を結ぶ．

3.1　異遷移同目的ドメイン

この節では，異遷移同目的ドメインについて説明を行う．2 章で説明したように，強化学習において転移学習を行うためには，エージェントと環境，タスクによって構成されるモデルを考慮する必要がある．

異遷移同目的ドメインとは，環境とエージェント，タスクにより構成されるドメインが，エージェントが観測する状態 S，エージェントがとりうる行動 A，エージェントに課せられたタスクの終了条件 G で表すことのできる転移学習のモデルの 1 つを指す．具体的な例としては，災害発生前と災害発生直後におけるそれぞれ

の環境において，同じ目的地へ荷物を運ぶケースがイメージしやすいだろう．このケースだと，災害発生前には道路として通行が可能であった場所が，災害発生直後には落下物が存在したり，地割れが起きていたりすることにより道路が寸断されているといった環境の変化が生じることがある．ただ，荷物を目的地に届けるというタスク自体は変化しない．このような状況下において，過去に覚えた道筋をたどることで目的地に着けるとは限らない．このとき，双方のドメインとして共通の要素は，エージェントは同一エージェントを考えるため，エージェントが観測する状態 S，エージェントがとりうる行動 A は固定される．また，エージェントに課せられたタスクに変化はないため，タスクの終了条件 G が固定される．従って，この異遷移同目的ドメイン D は $< S, A, G >$ と表すことができる．

なお，一般的に，マルコフ決定過程において，タスクの終了条件 G は特定の状態あるいは特定の状態遷移とされるため，陽に定義されないが，あえてここでは G を陽に定義している．その理由は，強化学習において G が異なることは，エージェントがとるべき方策が異なることにつながり，これにより，転移の仕方にもそれぞれのドメインにおいて適した手法を考える必要があると考えたためである．そのため，私はあえてゴール G を陽に定義することとした．

この節では，この異遷移同目的ドメインを対象とし，転移学習を用いた強化学習について検討を行う．

3.2 行動価値の推移

この節では，強化学習の学習中に更新される行動価値関数が転移学習における知識の選択する際，タスクの特徴とすることができるのかについて，検討を行う．強化学習では，学習により行動価値関数を最適な値にすることで知識を獲得している．つまり，エージェントの知識は状態と行動の価値の表により表されている．そこで，学習中の行動価値関数により獲得すべき行動規則を予測することを考える．

行動価値は，ゴールに近い状態ほど大きな値が与えられる．しかし，学習開始直後においては，最も大きな行動価値を持つ状態がゴールであるとは限らない．そのため，学習中の行動価値はゴールではない状態へ向かう可能性がある．つまり，学習中の行動価値関数では，ある状態における最大の行動価値が，選択すべき行動であると判断することが難しいと考えられる．

簡単な実験（本章の後半で行う実験と同等の実験）を行い，ある状態の行動価

値関数の推移を観測した．その結果を表 3.1 に示す．

表 3.1: 学習中における行動価値の推移

	行動 1	行動 2	行動 3	行動 4
学習序盤 (20[episode])	-2.7	0.4	1.3	-3.0
学習中盤 (50[episode])	-2.7	0.8	-2.6	-3.0
学習終盤 (70[episode])	-5.0	-0.9	-5.2	-4.6
タスク達成 (90[episode])	-7.1	-4.6	4.2	-6.7

表 3.1 において，この状態における最適な行動は行動 3 である．この 3.1 より，学習途中の各時点において，必ずしも行動 3 の行動価値が最も高いわけではないことが読み取れる．これは，学習途中の行動価値関数から，現在のタスクにおいて最適な行動が何であるかを推測することが困難であることが分かる．また，行動 1 や行動 4 のように常に低い行動価値を示すものが見られた．これらの行動は，負の報酬を与えられる行動であり，負の報酬が与えられる行動については，学習序盤から終盤まで一貫して行動価値が低いことが分かった．しかし，学習中盤から終盤にかけては行動 3 の行動価値も低くなっており，行動価値の推移から現在のタスクに適した行動を推測することは難しいといえる．

この傾向は他の状態，他の学習過程でも多く見られたことから，本研究では，別の方法により，転移させる知識の選択を試みる．

3.3 禁止行動の提案

前節では，行動価値の推移を観測することにより，ある状態における適切な行動の推測は困難であることを述べた．このことをふまえ本節では，学習中に知識の選択を行うのに使える手がかりについて検討を行う．

強化学習では，多くの場合，タスクの失敗を引き起こす行動に対して，大きな負の報酬が与えられる．これは，エージェントの故障や学習時間の増加を防ぐためである．報酬は学習過程においてかならずエージェントが観測しているため，学習中に情報を得ることができる．この大きな負の報酬を観測することで，エージェントにとって不都合な行動（以降，禁止行動とよぶ）を検知することが可能である．

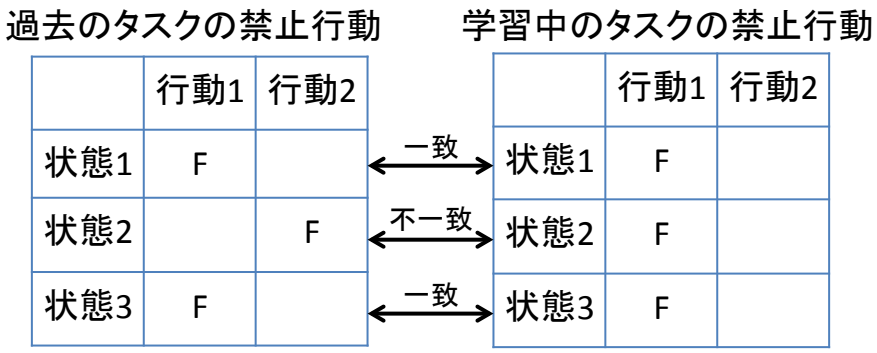


図 3.1: 禁止行動の比較

そこで，この禁止行動を学習過程で収集し，これを転移させる知識の選択に利用することを提案する．具体的な手順としては，

1. 禁止行動を学習中に獲られる報酬により検知する，
2. 各状態ごとに禁止行動のリスト（禁止行動規則）を作成する，
3. 各状態ごとに，それぞれの過去のタスクと現在のタスクの禁止行動を比較し，禁止行動の一致数をカウントする (図 3.1)，
4. 式 3.1 に従い，禁止行動規則の一致率を算出する，
5. 禁止行動規則の一致率が最も高く，しきい値 θ を超える知識を転移させる知識とする，

このように行われる．

$$\text{信頼度} : C = \frac{\text{禁止行動の一致数}}{\text{全状態数}} \tag{3.1}$$

過去のタスクにおける禁止行動規則と現在のタスクにおける禁止行動規則とを比較し，禁止行動規則の一致率が高い知識を転移させる知識とする．禁止行動規則の一致率が高い知識ほど現在のタスクでとるべきではない行動を多く含む．それとは，別に現在のタスクの達成に必要な行動が同じである可能性が高くなることが予想される．ただし，禁止行動規則の一致率が低い場合は，転移すべきではない知識であると思われるため，あるしきい値 θ よりも上回る知識がない場合は知識の転移を行わないようにする．

学習中に転移学習が発生するため，転移すべきではない知識も選ばれる可能性ある．そのため，この知識の選択手法では，現在転移させると選んでいる知識より，禁止行動規則の一致率が高い知識が見つかった場合には，転移させる知識が変更できるようにした．これにより，しきい値を低く設定して，転移すべきではない知識が選択されたとしても，より学習の高速化に貢献する知識が発見された場合には，より良い知識が転移されることになる．

これにより，これまでの転移学習を用いた強化学習のフローチャート (図 3.2) では，強化学習と転移学習とが別々となっており，一度転移させる知識を選んでしまった後は，知識を変更することができなかった．しかし，禁止行動規則を用いる手法では，図 3.3 のようになり，強化学習の流れの中で，転移学習が行えるため，転移させる知識は何度でも変更することができる．

次節において，禁止行動規則を用いた知識の選択手法により，学習の高速化につながる知識が選択されるのかを検証する．

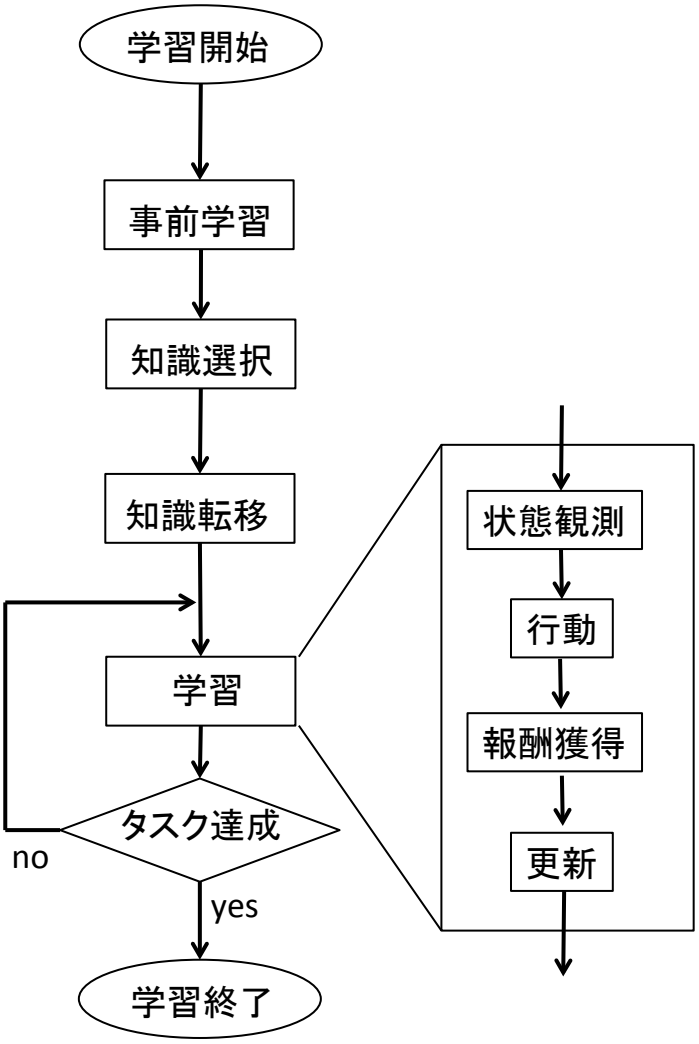


図 3.2: これまでの転移学習を用いた強化学習のフローチャート

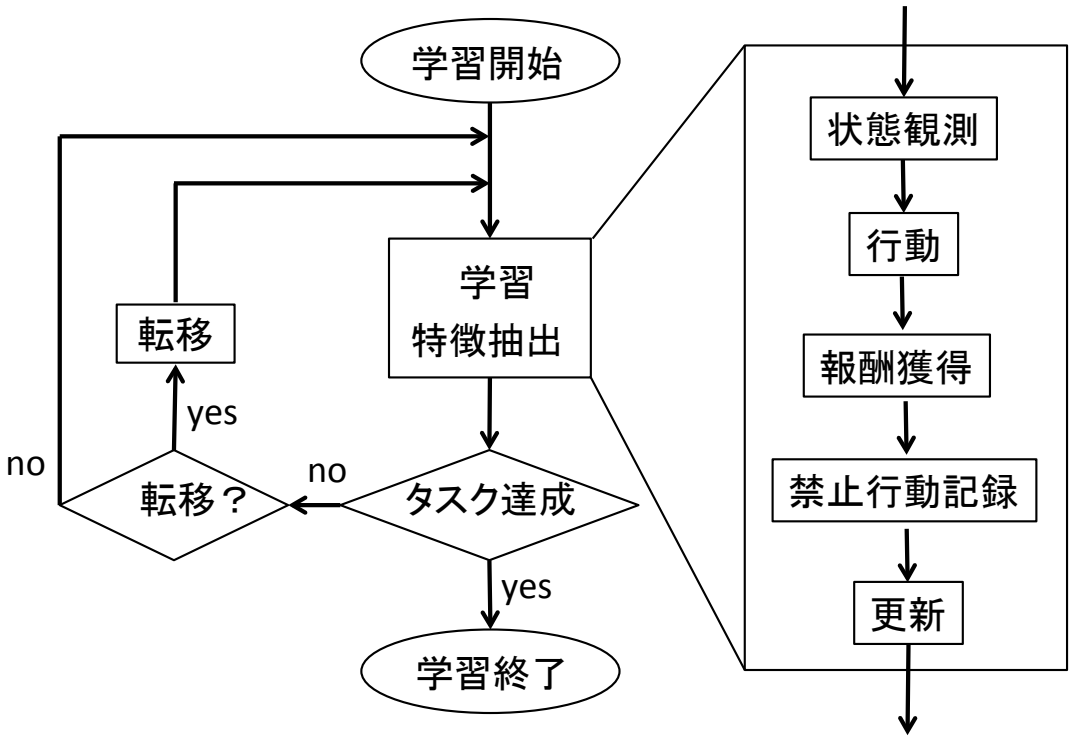


図 3.3: 提案する転移学習を用いた強化学習のフローチャート

3.4 禁止行動規則により選ばれた知識の検証

この節では，禁止行動規則を用いた知識の選択手法により，学習の高速化につながる知識が選択できるのかを検証する．

3.4.1 実験環境

実験には，エージェントは現在の位置のみを観測し，上下左右のいずれかの行動ができるエージェントに，簡単な迷路をスタートからゴールまで走破させるタスクを与えた（図 3.4）．迷路は，7×7 の格子からなり，白マスは通過してもよい通路，黒マスが侵入できない穴を意味している．図 3.4 の S はスタート，G はゴールをさしている．

過去のタスクとして図 3.5 に示すような迷路を用意し，それぞれのタスクにおいてエージェントは十分な学習を行い，それぞれのタスクを達成するための知識（行動価値関数，状態価値関数，禁止行動規則）を保持している．

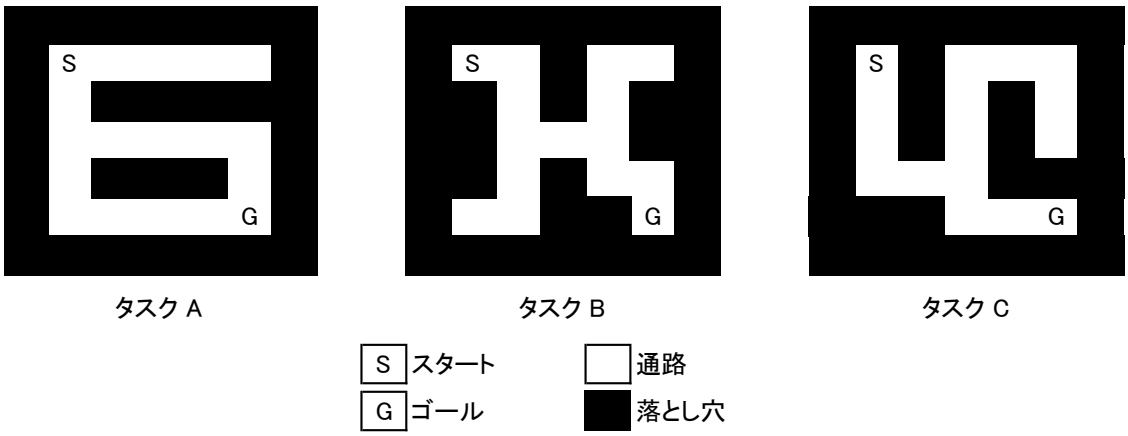


図 3.4: 簡単な迷路問題

なお，実験の条件は以下の通りに定めた．報酬 r は，壁にぶつかる行動に対して，$r = -50$，ゴールにたどり着いたら，$r = 100$，上下左右のいずれかの行動を 100 ステップ行うごとに $r = -25$ を与えた．また，アクタークリティックの学習パラメータは，割引率 $\gamma = 0.95$，学習率 $\alpha = 0.05$，ステップサイズパラメータ $\beta = 0.05$ とし，学習中の行動選択には $\epsilon-$greedy 法を使用し，$\epsilon = 0.05$ とした．

学習済みの知識には，学習によって得た行動価値関数だけでなく，禁止行動規則が保持されているものとする．転移させるかさせないかの判断基準として $\theta = 0.2$ 以上となる過去の知識を転移させる知識の候補とし，その中で最も高い禁止行動規則の一致率をもった知識を転移させる知識とした．これらを，100[trials] 行い，1[trial] は 2,000[episodes] を上限とし，それ以下の [episodes] で学習が完了できるかどうかとした．1[episode] あたり，100[step] の行動が行える．

それぞれのタスクにおいて，選択されるべき知識について述べる．

タスク A データ ID 4，データ ID 7，データ ID 10 の学習達成度が高く，選択されるべき知識である，

タスク B データ ID 21 のみが学習達成度が高く，選択されるべき知識である，

タスク C 特に選択されるべき知識はない，

となるように設定した．なお，学習達成度とは，現在のタスクにおいてとるべき行動が過去のタスクにおいてもとるべき行動となっている状態の比率を指してい

る．この数値が高い程，エージェントがとるべき行動が多いため，学習の高速化につながる．

3.4.2　実験結果と考察

転移させる知識の選択法として提案した手法により選ばれた知識が転移すべき知識であるかどうかを調べる実験を行った．表 3.2 にその結果を示す．ここで，データ ID は，図 3.5 のそれぞれのタスクにつけた番号に対応した知識を保持している．選択割合は，最終的に 100[trials] 中，この知識を転移すると決定した割合を意味している．また，学習達成度は現在のタスクの状態と行うべき行動が現在のタスクと過去のタスクとどれだけ同じであるかの比をとったものである．学習達成度により，過去の知識がどれだけ現在のタスクと一致しているかを表している．平均学習達成度は，データベース内の知識すべての学習達成度の平均である．平均学習達成度より高い知識を選んでいることは，学習の高速化につながらない知識を選別していないことを指しており，学習の高速化につながりやすいことを意味する．

表 3.2 を順に見ていく．まずタスク A では，選択された知識がデータ ID10，データ ID4，データ ID7 と学習達成度が高く，選択されるべき知識が上位 3 位を占めていることが分かる．タスク A において，知識の選択が行われたのは，総試行の選択割合としては 0.76 となっており，そのうち 0.64 は転移すべき知識が選択された．これはつまり，転移学習が行われたうち 84%が転移すべき知識を選択できたことを意味している．次にタスク B では，選択された知識がデータ ID21 のみとなった．これは，転移学習が行われたとき，必ず転移させるべき知識が選ばれていることを指している．そしてタスク C では，学習達成度が高くなる知識をデータベースに作らなかった．しかし，知識の選択が行われた割合が 0.61 となり，選ばれた知識の内 0.59 は平均学習達成度を超える知識が選ばれた．

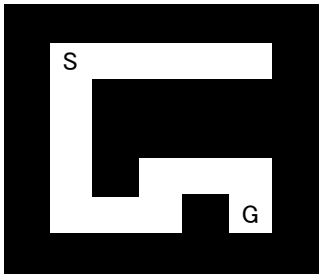
以上より，全体的に禁止行動規則を利用した知識の選択方法は，平均学習達成度以上の知識が選ばれていることがわかる．タスク C で示されるように，一部では学習達成度が低い知識が選ばれてしまっているが，選択割合は低い．これは，転移学習が行われやすいようしきい値 θ を低くしたため，学習達成度が低い知識も選ばれてしまったことが考えられる．このため，タスク C においては，特に選択すべき知識を用意しなかったが，転移学習が 61%も行われてしまった．

以上より，禁止行動規則を用いて転移させる知識を選ぶ手法は，知識データベースの中で比較的学習達成度の高い知識を選ぶことに有効であるといえる．しかし，

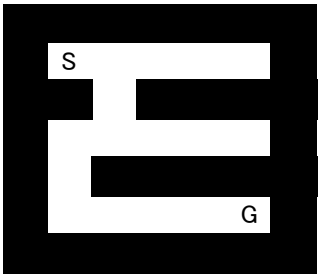
特に選択すべき知識がデータベースにない場合でも，転移学習が行われてしまうと結論づけられる．

表 3.2: 実験 1 で選択された知識の学習達成度

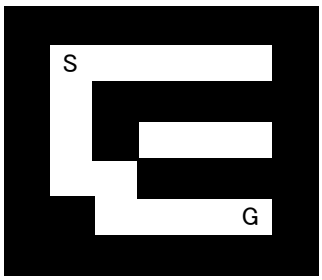
タスク	順位	データ ID	選択割合	学習達成度	平均学習達成度
A	1	10	0.52	0.56	0.43
	2	4	0.07	0.65	
	3	7	0.05	0.76	
B	1	21	0.33	0.79	0.21
	2	-	-	-	
	3	-	-	-	
C	1	15	0.54	0.33	0.27
	2	19	0.05	0.33	
	3	14	0.02	0.07	



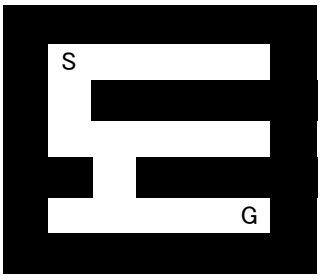
データ ID 1



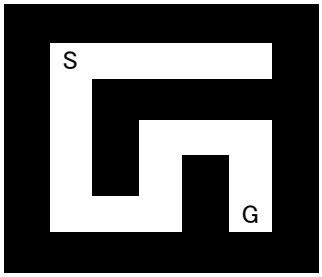
データ ID 2



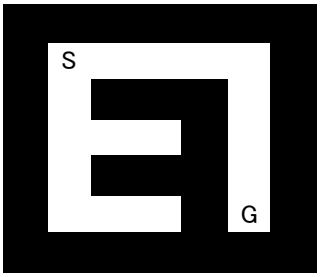
データ ID 3



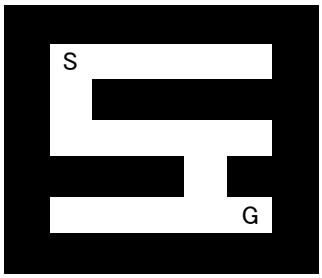
データ ID 4(タスク A に近い)



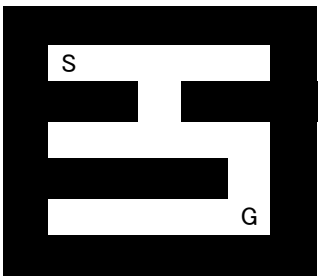
データ ID 5



データ ID 6

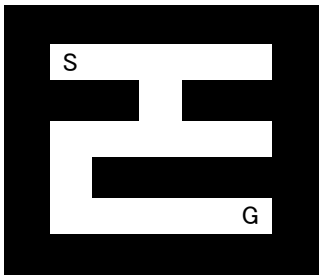


データ ID 7(タスク A に近い)

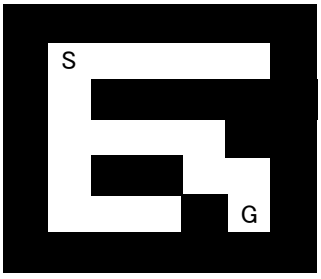


データ ID 8

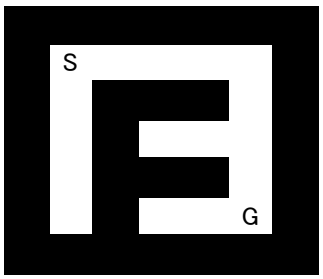
図 3.5: 簡単な迷路問題により得られた知識



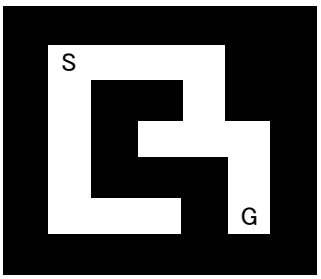
データ ID 9



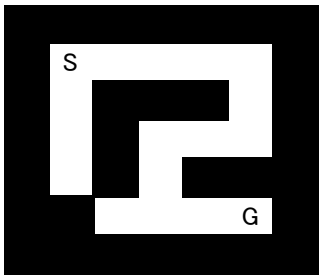
データ ID 10(タスク A に近い)



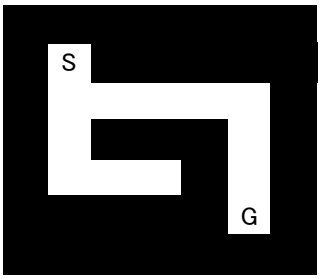
データ ID 11



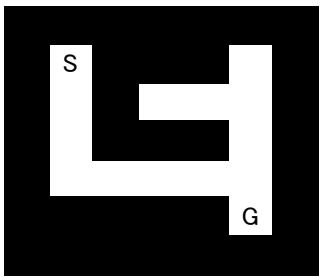
データ ID 12



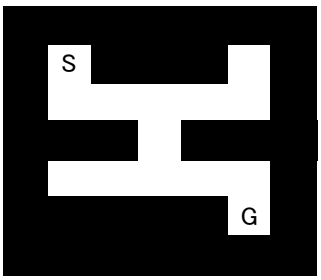
データ ID 13



データ ID 14

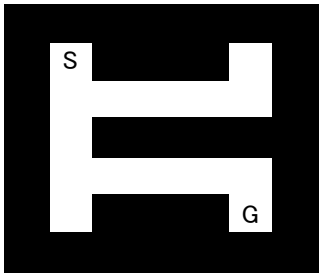


データ ID 15

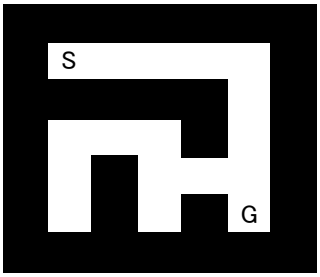


データ ID 16

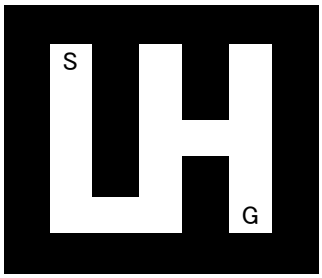
図 3.6: 簡単な迷路問題により得られた知識 (続き)



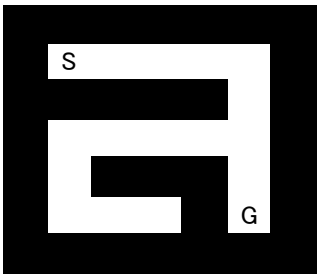
データ ID 17



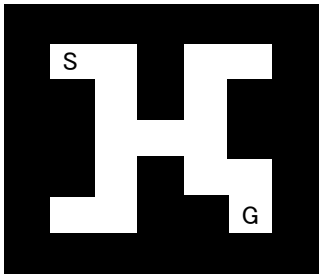
データ ID 18



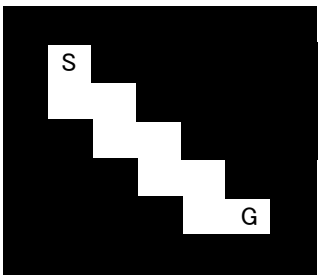
データ ID 19



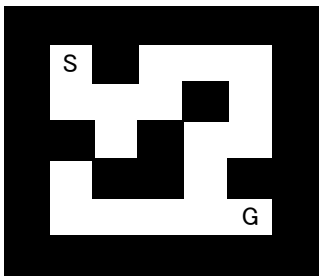
データ ID 20



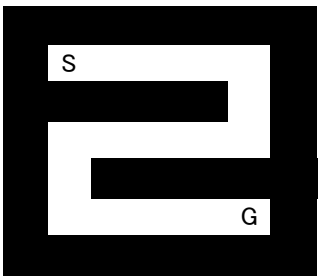
データ ID 21(タスク B に近い)



データ ID 22



データ ID 23



データ ID 24

図 3.7: 簡単な迷路問題により得られた知識 (続き)

3.5　知識の統合

前節の実験により，禁止行動規則を用いて知識を選択することで，学習を高速化できそうな知識をデータベースから選別することが可能であることが分かった．しかし，データベースに学習の高速化につながる知識がない場合にも転移学習が行われてしまうことも分かった．そのため，本節では，過去の知識を利用しても，探索が行われやすい知識の利用法について検討する．

3.5.1　加重平均

転移学習により，エージェントは過去の知識を利用することが可能である．しかし，現在の学習により得た知識も重要であり，双方の知識は現在の学習において積極的に利用されるべきである．強化学習における転移学習では，知識が行動価値関数として表され，これらは数値情報である．過去の知識も現在の知識も利用する手法が提案されている [71]．[71] は確率的に過去の行動価値関数に基づいた行動と現在の学習により得ている行動価値関数に基づいた行動とに切り替える手法である．なお，学習が進むにつれ，現在の行動価値関数に基づいた行動が選択されやすくなる．この手法では，徐々に行動選択確率が変遷するため，不要な試行が行われることが心配される．

そこで，行動価値関数が数値情報であることを利用して，過去の知識と現在の知識を統合して 1 つの行動価値関数を生成することを提案する．強化学習においては，行動価値関数や状態価値関数で与えられる．これらは，数値情報であるため，単純に足し合わせることが可能である．また，過去の知識と現在の知識の比重は必ずしも同じで良いとは限らない．前節での実験を受けると，過去の知識を信頼する場合も有れば，過去の知識をそれほど信頼してはいけないことも考える必要がある．そのため，ここでは，どちらの場合も考慮できるよう，加重平均をとることにする．具体的には，以下の式によって重み付き平均をとる．

$$p_{\mathrm{t}}(s_n, a_n) \leftarrow \frac{1}{2}\{(1-\zeta)p_{\mathrm{t}}(s_n, a_n) + \zeta p_{\mathrm{s}}(s_n, a_n)\}, \tag{3.2}$$

$$V_{\mathrm{t}}(s_n) \leftarrow \frac{1}{2}\{(1-\eta)V_{\mathrm{t}}(s_n) + \eta V_{\mathrm{s}}(s_n)\}. \tag{3.3}$$

ここで，ζ と η はそれぞれ重み係数を表し，添字 s は過去の知識であることを意味しており，添字 t は現在の知識であることを意味している．

加重平均をとることで，現在の行動価値と過去の行動価値とのバランスを調整できるため，転移させる知識の学習達成度が高い場合は，過去の行動価値関数を重視し，転移させる知識の学習達成度が低い場合は，現在の行動価値関数を重視することになる．ただし，学習達成度について現状では把握することが難しいため，学習達成度を推測するための議論が別途必要である．

3.5.2 Strategy R

前項で述べた過去の行動価値関数と現在の行動価値関数の加重平均をとることで，転移学習を用いた強化学習は実現できる．しかし，過去の知識を転移して即座に現在のタスクが達成できることはまれである．そのため，過去の知識を転移した後にも現在のタスクを達成するよう行動価値関数の調整が必要となる．そのとき，過去の知識により行動価値関数の更新が思うように進まないことが考えられる．これは，禁止行動規則による選択された知識の学習達成度は必ずしも高いわけではなく，これまでの実験でも，学習達成度が100%ではなっていないことから，転移させる知識の中には，転移すべきではない情報が含まれている可能性は否定できない．ここで，間違った情報とは，過去の知識では，エージェントがとるべき行動であるが，現在のタスクにおいては，エージェントがとってはいけない行動のことを指している．逆のケースも考えられるが，これまでの実験からも分かるように，現在のタスクにおいてエージェントがとるべき行動を推定することは難しいため，本論文では，上記のように間違った情報を定義する．間違った情報を転移することは，エージェントに無駄な行動をさせ，学習の進行を阻害する要因となる．そのため，間違った情報はできる限り転移させない工夫について検討を行う．

学習中に知識の転移を行うためには，現在のタスクについての情報がある程度集まっていることを意味しており，過去の知識についてもデータベースに保持されている．そこで，これらの知識から，明らかに間違った情報を見つけることを考える．本研究では，学習中に転移学習を行うため，過去の知識にも現在の知識にも前述した禁止行動規則を保持している．そこで，この禁止行動規則を利用すると，図3.8のような状況を認知できる．図3.8はある状態においてエージェントの禁止行動を描いたものである．図中の×で示される行動が禁止行動，過去の状態において○で示される行動がそのときエージェントがとるべき行動を指している．図3.8では，現在の状態においては右に進む行動が禁止行動であることが経験

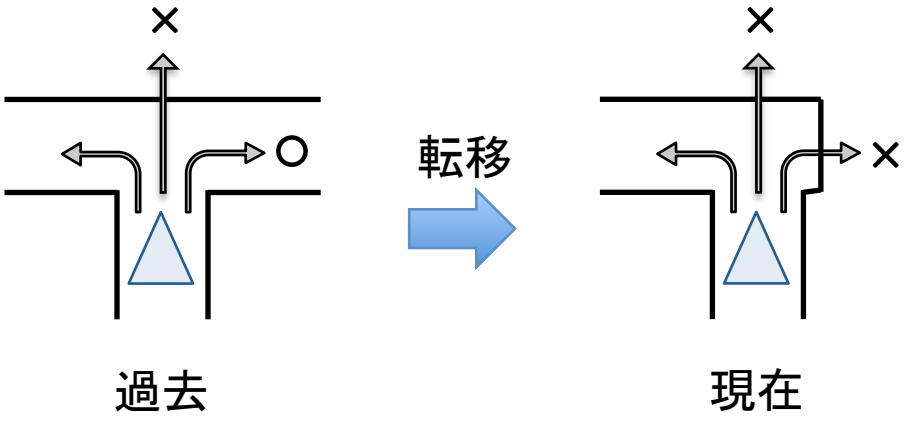


図 3.8: 間違った情報の転移例

により分かっている．このとき，転移により選択する行動は明らかに間違った情報であり，禁止行動を利用することで判断することができると考えられる．

このように，ある状態において，過去の知識によりとるべき行動と現在のタスクにおける禁止行動に矛盾が生じる場合，その状態の情報は明らかに間違った情報であると考えられる．こういった矛盾のある情報については，宮崎らは [42] において，罰を回避することでエージェントは合理的政策を得ることができると述べている．つまり，間違った情報以外の情報の中には，エージェントがとるべき行動が含まれていると考えられる．

過去の知識においてとるべき行動と現在のタスクにおける禁止行動を比較して転移する手法を Strategy R とよぶこととし，具体的には，式 3.4 によって実現できると考えられる

$$\{p(s_n, a_n) | \forall s_n \in \text{equivalent states}, \forall a_n \in A\}. \tag{3.4}$$

Strategy R はエージェントに，過去の知識を吟味し間違った転移をさせないようにすることを目的としており，直接的に行動選択に関係する行動価値関数にのみ適用することが望ましいと考えられる．

3.5.3 Strategy P

学習の早い段階で転移学習を起こすと前項で述べた Strategy R はほとんど過去の知識を利用できない不安がある．この項では，直接的に行動選択に影響を与えな

い状態価値関数を利用し，Strategy R の消極的な側面の改善を図る Strategy P を提案する．行動価値関数は，状態価値関数により価値が更新される．そのため，大きな状態価値を持つ状態に遷移するよう行動価値関数は更新されやすい．Strategy P は，この性質を利用し，大きな状態価値を転移させることにより行動価値関数の更新を誘導し，Strategy R の消極的な転移を補うことを期待している．この Strategy P は式 3.5 の形で示される

$$\{V(s_n)|V(s_n) > 0, \forall s_n \in S\}. \tag{3.5}$$

3.6 提案アルゴリズム

この節では，ここまでに提案された転移学習を用いた強化学習の手法をまとめ，アルゴリズムとして図 3.9 に示す．ここで，P は行動価値関数，V は状態価値関数を表し，F は禁止行動規則，L はデータベース，C は禁止行動規則の一致率をそれぞれ表している．また，添字 p は現在選択していることを指し，添字 e は最も効果的であると判断された知識，添字 d はデータベースの中にある知識をそれぞれ指している．

次節において，本章で提案した転移学習を用いた強化学習における学習の高速化について簡単な実験により検証を行う．

3.7 異遷移同目的ドメインにおける転移学習による学習の高速化の検証

この節では，本章で提案した転移学習を用いた強化学習により，現在のタスクを達成する行動規則を獲得するまでの時間が削減されるのかについて検証を行う．実験環境については，3.4 と同じ環境のため，ここでは説明を割愛する．実験では，加重平均による転移学習の効果と条件付き加重平均による転移学習の効果についてそれぞれ検証を行う．

3.7.1 加重平均による転移学習の効果

ここでは，異遷移同目的ドメイン間において加重平均を用いた転移学習による学習の高速化について検証を行う．

//初期設定

initialize parameters P and V.

$\phi \rightarrow$ forbidden rule set F

( ) $\rightarrow$ the latest transferred item $(P_{\mathrm{p}}, V_{\mathrm{p}}, F_{\mathrm{p}})$.

`while(` agent does not satisfy termination conditions `) {`

observe state $s \in S$.

decide action $a \in A$.

receive reward r.

// 禁止行動規則の更新

`if(` a is a forbidden action `) {`

add (s, a) into F.}

// 過去の知識の選択

( ) $\rightarrow$ the most effective item $(P_{\mathrm{e}}, V_{\mathrm{e}}, F_{\mathrm{e}})$.

$0 \rightarrow$ the highest concordance rate C_{e}.

// 転移させる知識の更新

`foreach(` $(P_{\mathrm{d}}, V_{\mathrm{d}}, F_{\mathrm{d}})$ `in` database L `) {`

concordance rate for F_{d} to $F \rightarrow C$.

`if(` $C > C_{\mathrm{e}}$ `) {`

$(P_{\mathrm{d}}, V_{\mathrm{d}}, F_{\mathrm{d}}) \rightarrow (P_{\mathrm{e}}, V_{\mathrm{e}}, F_{\mathrm{e}})$.

$C \rightarrow C_{\mathrm{e}}$.$\}\}$

`if(` $C_{\mathrm{e}} > \theta$ `&&` $(P_{\mathrm{e}}, V_{\mathrm{e}}, F_{\mathrm{e}})$ `!=` $(P_{\mathrm{p}}, V_{\mathrm{p}}, F_{\mathrm{p}})$ `) {`

// 知識の転移

extract appropriate parameters.

action preferences $\{p'_{\mathrm{e}}\}$ by equation (3.4) from P_{e}.

state values $\{v'_{\mathrm{e}}\}$ by equation (3.5) from V_{e}.

merge $\{p'_{\mathrm{e}}\}$ into P according to equation (3.2).

merge $\{v'_{\mathrm{e}}\}$ into V according to equation (3.3).

$(P_{\mathrm{e}}, V_{\mathrm{e}}, F_{\mathrm{e}}) \rightarrow (P_{\mathrm{p}}, V_{\mathrm{p}}, F_{\mathrm{p}})$.}

`else {`

update P and V (actor-critic method).}

}

図 3.9: 異遷移同目的ドメインにおける転移学習の流れ

表 3.3: 加重平均を用いた転移学習における平均学習回数

	Original	加重平均		
		P	V	P-V
Ω_A	250.4 (38)	150.4 (29)	259.0 (20)	184.0 (17)
Ω_B	231.1 (66)	202.9 (25)	209.0 (17)	190.3 (19)
Ω_C	281.1 (147)	194.7 (103)	300.4 (130)	281.7 (142)

強化学習ではそれぞれの加重平均をとった場合，以下の組み合わせが考えられる．

- 行動価値関数のみ加重平均をとる．
- 状態価値関数のみ加重平均をとる．
- 行動価値関数と状態価値関数の加重平均をとる．

これと通常の強化学習 (アクター・クリティック) による学習回数の比較を表 3.3 に示す．表 3.3 はそれぞれのタスクにおける平均学習回数を示している．() は各タスクにおける学習失敗回数を示している．表中 P は行動価値関数のみを転移した場合，V は状態価値関数のみを転移した場合，P-V は行動価値関数と状態価値関数を転移させた場合を表している．また，灰色のセルは，通常の強化学習手法と比べ，T 検定により有意な差が得られたことを表している ($p < 0.01$)．

この結果では，強化学習の行動価値関数と状態価値関数は加重平均をとることで，学習回数の削減あるいは，同程度の学習回数ですんでいることがいえる．つまり，異遷移同目的タスクという条件下で，強化学習における転移学習は，各パラメータの加重平均をとることで，学習の高速化につながるといえる．

3.7.2 条件付き加重平均を用いた転移学習の効果

重み付き平均にさらに条件を加えた転移手法について検討する．この条件は，4 章で説明したように，禁止行動が一致する信頼性の高い情報の転移と行動選択に関連する価値の高い情報のそれぞれについて転移する手法である．2 つの条件と 2 つのパラメータにより，考えられる組合せは以下の 9 通りである．

- 通常の強化学習

- Strategy R による行動価値のみ転移
- Strategy R による行動価値と Strategy R による状態価値を転移
- Strategy R による状態価値のみ転移
- Strategy R による行動価値と Strategy P による状態価値を転移
- Strategy P による行動価値のみ転移
- Strategy P による行動価値と Strategy R による状態価値を転移
- Strategy P による状態価値のみ転移
- Strategy P による行動価値と Strategy P による状態価値を転移

この 9 通りについて検証を行う．この実験結果を表 3.4 に示す．

表 3.4: 条件付き加重平均を用いた転移学習の効果

(a) Task Ω_A

V \ P	Strategy R	Strategy P	No Transfer
Strategy R	221.2 (39)	212.8 (21)	286.7 (31)
Strategy P	209.6 (11)	204.1 (12)	247.1 (15)
No Transfer	208.2 (31)	196.7 (18)	250.4 (38)

(b) Task Ω_B

V \ P	Strategy R	Strategy P	No Transfer
Strategy R	203.1 (31)	194.9 (29)	217.9 (49)
Strategy P	192.1 (17)	205.5 (20)	213.4 (19)
No Transfer	210.0 (34)	193.8 (29)	231.1 (66)

(c) Task Ω_C

V \ P	Strategy R	Strategy P	No Transfer
Strategy R	196.7 (56)	291.7 (175)	284.1 (165)
Strategy P	195.4 (64)	288.0 (119)	277.9 (131)
No Transfer	183.1 (54)	204.6 (84)	281.1 (147)

表 3.4 はそれぞれのタスクにおける学習回数の平均である．() はタスクの学習失敗回数を示している．表中の灰色のセルは，中段左の実験結果に対して，T 検定により有意な差が認められたものである ($p < 0.01$)．

まず，Strategy R について考察を行う．Strategy R は転移に注意を要する行動価値関数に向けて提案したものである．表 3.4 の (a)，(b)，(c) の下段左をみるとそれぞれにおいて，通常の強化学習と比べ，学習回数が削減される傾向にある．また，上段右をみると Strategy R を状態価値関数に適用しても，通常の強化学習と比べ，学習回数の削減にはつながっておらず，場合によっては，学習の進行を阻害している．つまり，Strategy R を行動価値関数に用いることにより，学習回数の削減につながると推測できる．

次に，Strategy P について考察を行う．Strategy P はパラメータの更新にあたり，よく用いられそうな値のみを転移する手法となる．これは，エージェントの行動に直接影響を及ぼさない状態価値関数に向けて提案したものである．表 3.4 各タスクにおける下段中をみると，行動価値関数に Strategy P を用いることで，タスクの失敗回数を削減できていることがわかる．これは，Strategy P で行動価値関数を転移した結果，エージェントのタスクを失敗する行動がすくなくなったということを意味する．また，中段右の状態価値関数に Strategy P を用いた場合，学習回数の削減にはいたっていないが，失敗回数の削減する傾向が見られる．これは，Strategy P を用いた状態価値関数には行動価値関数の更新に作用し，エージェントのタスクを失敗する行動を少なくしたと考えられる．

最後にそれぞれの価値関数に向けた Strategy を適用した場合（中段左）について考える．提案手法は，通常の強化学習と比べ，学習回数を少なくとも 16%削減することができ，失敗数を 55%削減することができた．それぞれのタスクで見ていけば，提案手法より学習の高速化をしている手法が見られるが，どのタスクにおいても学習の高速化をしている手法は，提案手法であるといえる．

これにより，異遷移同目的タスクにおいて禁止行動規則を用いた知識の選択と条件付き加重平均による知識の転移法についての提案手法の有効性が示されたといえる．

3.8 結び

本章では，異遷移同目的ドメイン間における転移学習を用いた強化学習について検討を行った．まず，知識データベースから知識を選択する手法として，エージェントが受け取る大きな負の報酬に着目し，禁止行動による，それぞれの過去の知識との類似性を測る手法を提案した．禁止行動規則を用いた過去の知識の選択手法では，データベース内にある知識から比較的学習達成度が高い知識が選択されていることを簡単な実験を通して確認を行った．

次に，禁止行動規則の一致率により選択された知識を利用するにあたり，間違った情報の転移が行われにくくなるよう条件付き加重平均について提案を行った．実験では，加重平均を用いた転移学習では，学習回数の削減ができる傾向を確認し，さらに条件付き加重平均を用いることで，加重平均を用いた転移学習より安定して学習の高速化につながることを確認した．ちなみに，実験では提案した転移学習により，タスクの失敗回数を 55%削減することができ，タスクの学習回数を 16%以上削減したことが確認された．

第4章　同遷移異目的ドメインにおける転移学習

本章では，同遷移異目的ドメインに着目し，このドメインにおける転移学習の手法について検討を行う．

まず，4.1 ではこの章で対象とする同遷移異目的ドメインについて説明する．4.2 では，禁止行動規則による知識の選択がこのドメインにおいても有効であるかを簡単に検証を行う．4.3 では，同遷移異目的ドメインにおいても転移学習が有効にはたらく知識の選択手法と知識の参照，および探索ついて検討を行う．4.4 にて，同遷移異目的間ドメインにおける転移学習を用いた強化学習の一連の提案をアルゴリズムとして示す．4.5 にて，学習の高速化につながりそうな知識が選ばれているのか，また学習の高速化が行えているのかを簡単に検証する．最後に，4.6 で本章を結ぶ．

4.1　同遷移異目的ドメイン

この節では，同遷移異目的ドメインについて説明を行う．2 章で説明したように，転移学習を強化学習に適用するためには，エージェントと環境，タスクによって構成されるモデルを考慮する必要がある．

同遷移異目的ドメインとは，ドメインの構成要素として，エージェントが観測する状態 S，エージェントがとりうる行動 A，エージェントが遷移する状態遷移確率 T により表すことのできる転移学習のモデルの 1 つを指す．具体的な例としては，地図のある地点 A からある地点 B に向かうとき，ある地点 B からある地点 A へと同じ経路辿り戻る場合を考える．地点 A から地点 B にいくタスクと地点 B から地点 A に向かうタスクでは，周囲の環境に違いはなく，エージェントのとるべき行動手順が，行きと帰りでは真逆となる．このような場合，エージェントの観測する状態 S，エージェントがとりうる行動 A，エージェントが遷移する状態遷移確率 T は同じ周囲の環境のため 3 つの要素が固定されることになる．また，タス

クの終了条件 G は与える目的が異なっていることから固定されず，これに伴って，獲得する報酬 R も変化する．このとき，ドメイン D は $< S,\ A,\ T >$ と表すことができ，このドメインのことを同遷移異目的ドメインと呼ぶこととする．

このドメインにおいては，状態遷移確率が変化しないため，禁止行動規則についてもタスク間での差が出ないことが予想され，前章で説明した禁止行動規則による知識の選別は難しいことが考えられる．

4.2 同遷移異目的ドメインにおける知識の選択方法の適用

同遷移異目的ドメインにおける転移学習では，前章で提案した知識の選択方法では，転移させるのに有効な知識を選び出すことが難しい．この節では，前章で提案した知識の選択方法を実際に実験に用いることで，その難しさについて説明を行う．

4.2.1 実験環境

本節の実験では，エージェントは現在の位置のみを観測し，上下左右のいずれかの行動をするエージェントを図 4.1 に示すようなタイルワールドのスタート (図中 S で示される地点) からゴール (図中 A，B，C で示されるいずれかの地点：それぞれを Task Ω_{A}，Task Ω_{B}，Task Ω_{C} と表記する) まで走破させるタスクを与えた．タイルワールドは，21×24 の格子からなり，白マスが通過できる通路，黒マスが侵入できない壁を表している．

また，学習済みの過去の知識として，図 4.2 を用意した．それぞれのタスクは図中 1 から 5 で示されるゴールまで走破する知識が保持されている．なお，データベース内の知識については，どの白マスからでもそれぞれのタスクを走破できる情報が保持されている．

なお，実験における強化学習のパラメータは以下のとおりに設定した．報酬 r はゴールにたどり着いた場合のみ $r = 1$ を与え，その他の場合については $r = 0$ とした．また，強化学習の手法としてはアクター·クリティックを使用し，アクター·クリティックの学習パラメータは，割引率 $\gamma = 0.95$，学習率 $\alpha = 0.05$，ステップ

サイズパラメータ $\beta = 0.05$ とし，行動の選択にはボルツマン選択を使用し，温度係数 $T = 1$ とした．

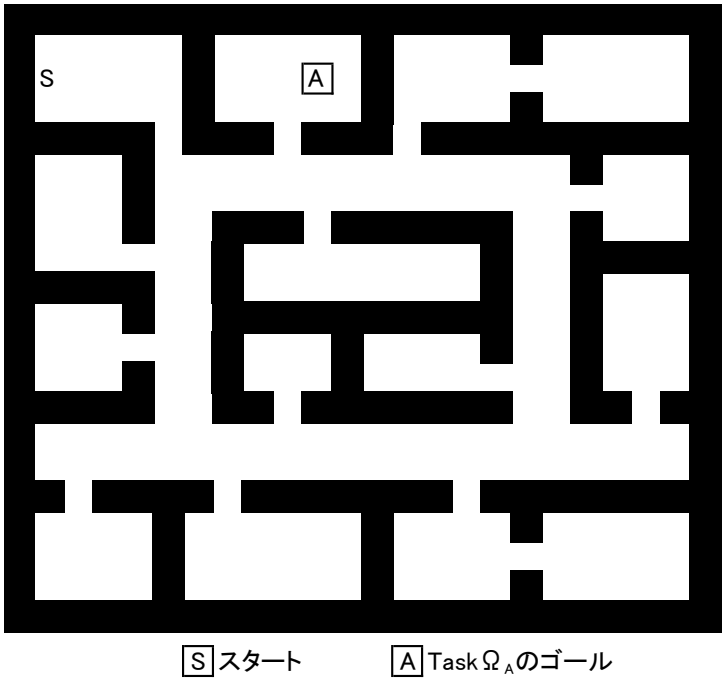
それぞれのタスクにおいて，選択されるべき知識は以下のとおりである．

Task Ω_{A}： データ ID Ω_1，データ ID Ω_2 が Task Ω_{A} のゴールに近く選択されるべき知識である．これは，複数の知識が転移学習を行うことで学習が高速化されることを意図して，ゴールを作成した．

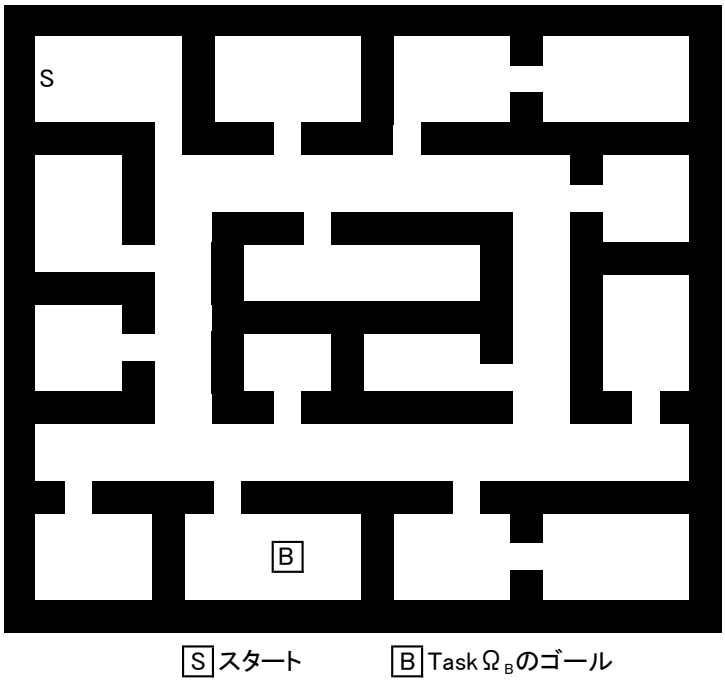
Task Ω_{B}： データ IDΩ_3 のみが Task Ω_{B} のゴールに近く選択されるべき知識である．これは，Task Ω_{B} ゴールは近いが，データ IDΩ_3 ゴールまでの経路からは少し外れた位置になるように，ゴールを設定した．

Task Ω_{C}： データ IDΩ_4 のみが Task Ω_{C} のゴールに近く選択されるべき知識である．これは，データ IDΩ_4 の経路上から近い位置になるよう Task Ω_{C} のゴールを設定した．

となるように設定した．学習済みの知識には，学習によって得た行動価値関数だけでなく，禁止行動規則も保持されているものとする．転移させるかさせないかの判断基準として $\theta = 0.2$ 以上となる過去の知識を転移させる知識の候補とし，その中で最も高い禁止行動規則の一致率をもった知識を転移させる知識とした．これらを，1000[trials] 行い，1[trial] は 100,000[episodes] を上限とし，それ以下の [episodes] で学習が完了できるかどうかとした．1[episode] あたり，100[step] の行動が行える．

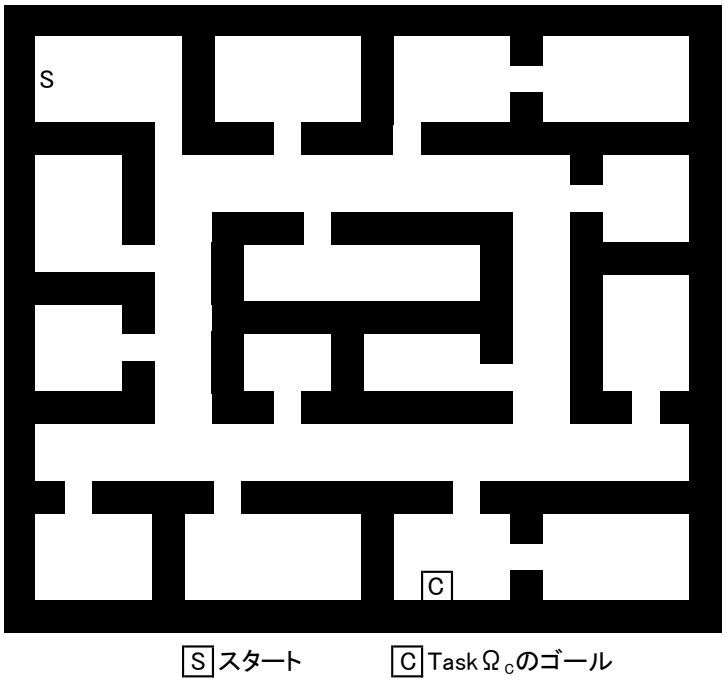


(a) Task Ω_{A}



(b) Task Ω_{B}

図 4.1: タイルワールド



(c) Task Ω_C

図 4.1: タイルワールド (続き)

4.2.2 実験結果と考察

表 4.1: 禁止行動規則による同遷移異目的ドメインにおける各知識の選択回数

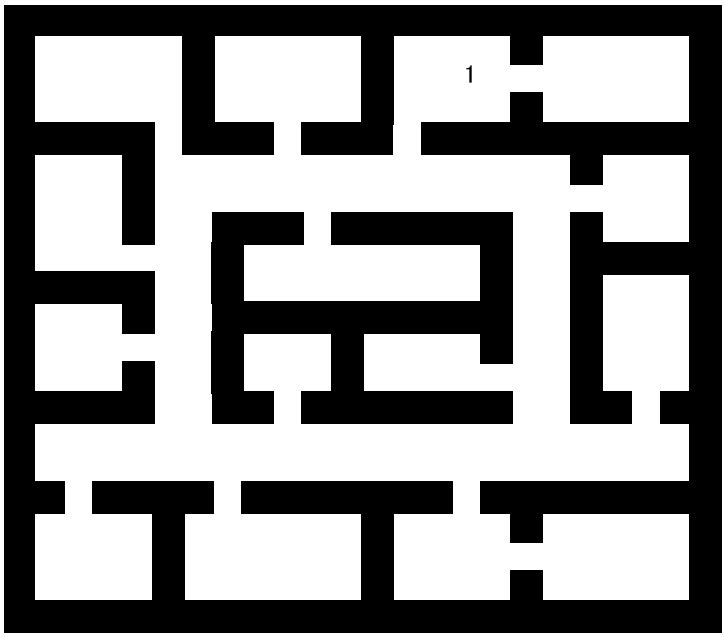
Task	データ ID Ω_1	データ ID Ω_2	データ ID Ω_3	データ ID Ω_4	データ ID Ω_5
Task Ω_A	414	1	231	368	94
Task Ω_B	4	1	6	2	987
Task Ω_C	3	3	2	7	985

異遷移同目的ドメインにおける転移学習の知識の選択方法を，同遷移異目的ドメインにおける転移学習の知識の選択方法として用いた結果を表 4.1 に示す．灰色のセルは過去の知識の中でも選ばれるべき知識を指している．

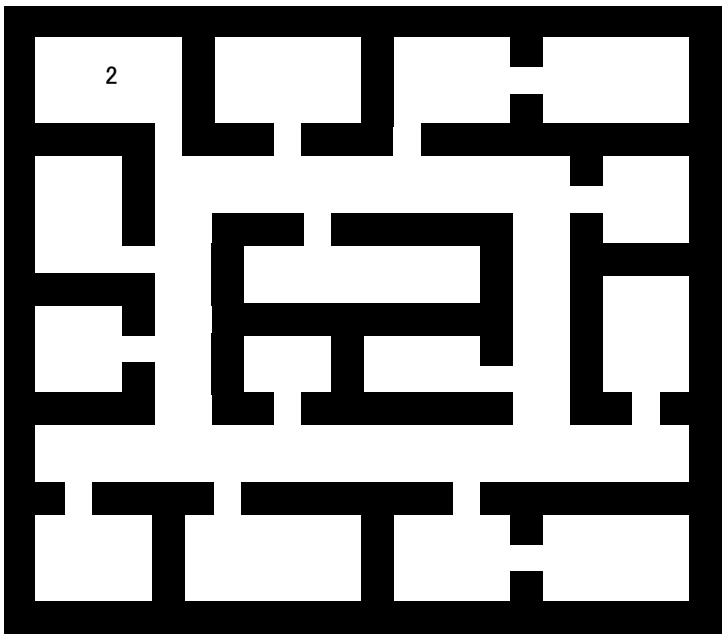
表 4.1 から分かるように，Task Ω_A では 42%程度，Task Ω_B では 0.6%，Task Ω_C では 0.7%と Task Ω_A でも 50%を超えないほど，選ばれるべき知識が選択できていない．

これは，異遷移同目的ドメインにおける転移学習の知識の選択方法では，禁止行動規則によりタスクの特徴を表しているが，同遷移異目的ドメインでは，禁止行動の違いが顕れないことを指している．なぜならば，禁止行動規則は状態遷移確率を別の形として捉えているに過ぎない．同遷移異目的ドメインではこの状態遷移確率が同じであるため，禁止行動規則によってそれぞれのタスクの違いを表すことが難しい．

ちなみに，表 4.1 では，同じ状態遷移確率のはずだが，かなり偏った結果が示されている．これはシミュレーション環境の実装の仕方の問題であり，議論に影響を及ぼすものではない．次節以降で，同遷移異目的ドメインにおけるそれぞれのタスクの違いを表す手法について検討を行う．

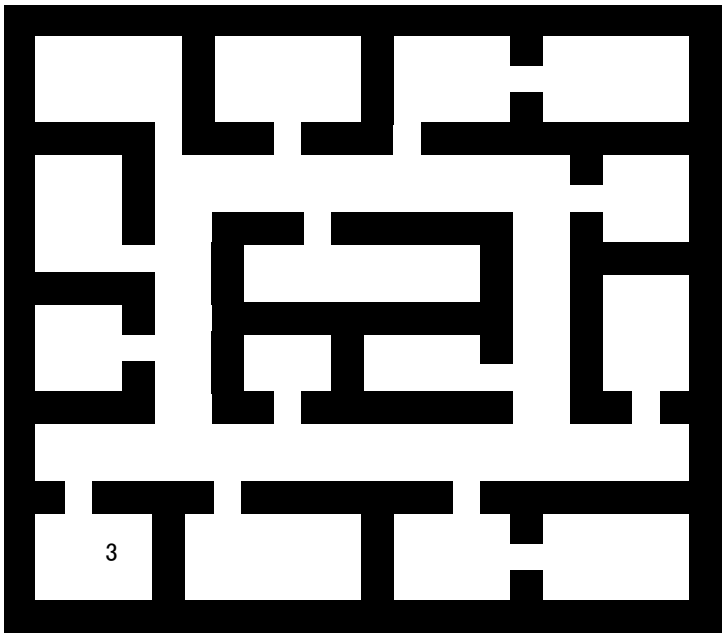


(1) データ ID Ω_1(Task Ω_A にゴールが近い)

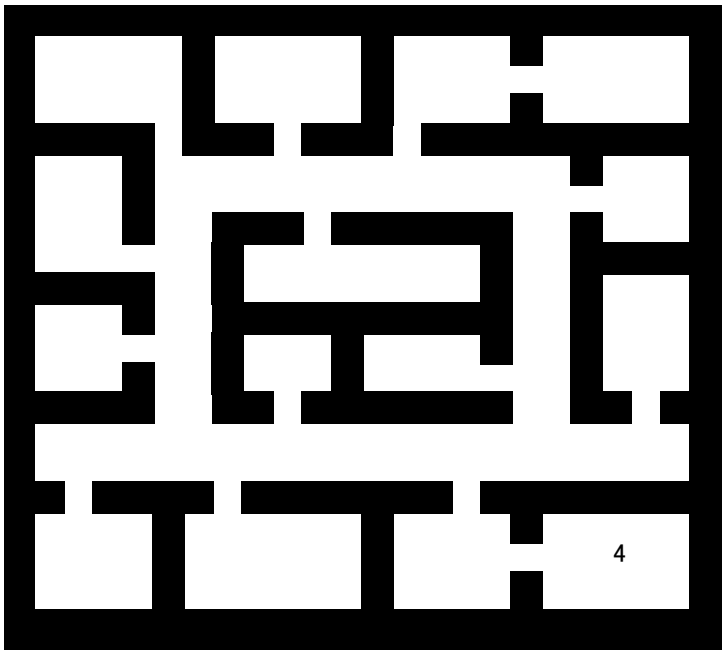


(2) データ ID Ω_2(Task Ω_A にゴールが近い)

図 4.2: 過去に学習したタイルワールド

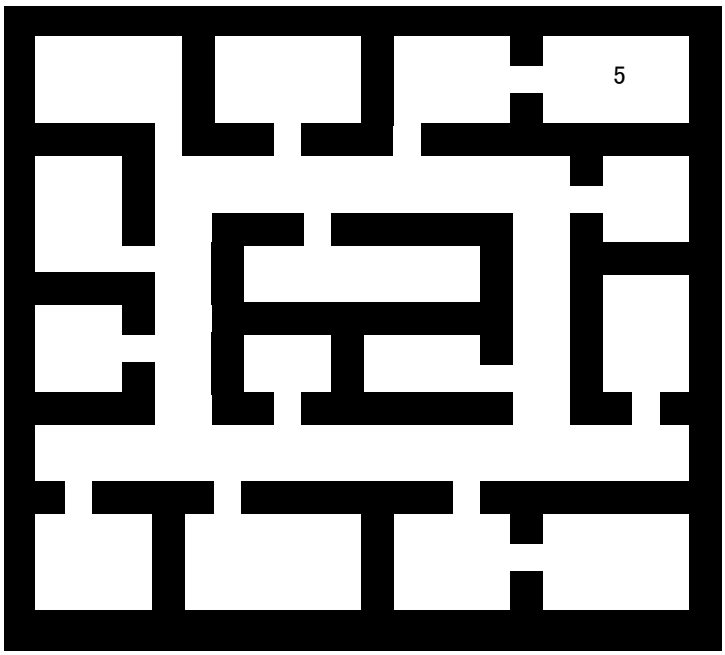


(3) データ ID Ω_3(Task Ω_B にゴールが近い)



(4) データ ID Ω_4(Task Ω_C にゴールが近い)

図 4.2: 過去に学習したタイルワールド (続き)



(5) Task Ω_5(どのタスクのゴールからも遠い)

図 4.2: 過去に学習したタイルワールド (続き)

4.3 同遷移異目的ドメインにおける転移学習の提案

この節では，同遷移異目的ドメインにおける転移学習の手法について検討を行う．

4.3.1 同遷移異目的ドメインにおける禁止行動規則による知識の選択方法の改善

同遷移異目的ドメインにおけるタスクの決定的な違いは，ゴールの違いである．周囲の環境は変化せず，ゴールが異なることがタスクの違いを表している．このドメインにおいて，転移学習がになうべき役割は，いかにゴールが近い知識を選ぶかである．しかし，ゴールが分かっているならば，転移学習を行う必要がほとんどない．なぜなら，ゴールが分かれば，あとは行動価値関数を強化するだけで良いのだから．となるとここでの転移学習の役割は，いかに多くの状態を探索できるかということにシフトすることになる．前章で提案した式 3.2 や式 3.3 は過去の知識と現在の知識を統合するため，知識の修正に少なからず時間をかけることになる．同遷移異目的ドメインの場合，状態遷移確率が同じであること，ゴールが異なっていることをふまえると，多くの過去の知識を使用し，多くの状態を探索させることが必要となる．

このドメインにおける従来の研究として [71] が挙げられる．[71] では，過去の知識を転移学習してみて，その結果現在の学習に利用すべきものかを判断している．しかし，彼らの手法では，すべての知識を一度は利用してみるため，データベースが大きくなるとそれだけ学習に時間がかかることになる．このとき，エージェントは，過去の行動価値関数と現在の行動価値関数とを確率的に使うようにし，学習が進むにつれて現在の行動価値関数を使用する頻度があがるように提案を行っている．

ここで一旦，禁止行動規則を用いた知識の選択方法を同遷移異目的ドメインにおける転移学習に用いた時の話に戻す．同遷移異目的ドメインにおいて，禁止行動規則を用いた知識の選択方法では，過去のタスクと現在のタスクにおいて状態遷移確率が同じであるため，過去のタスク間に差異が現れなかった．逆に捉えると，異遷移異目的ドメインにおける知識の選択方法において，禁止行動規則の一致率を変化させることで，過去のタスク間に転移すべきかすべきでないかの優劣を付けることが可能であると考えられる．このために，[71] を参考にして過去の知識を一度利用してみる必要がある．一度過去の知識を利用してみて，利用すべき

ならばその知識は禁止行動規則の一致率において積極的に利用するように，利用すべきではないのならば禁止行動規則の一致率において利用しないように調整を行うことで，これが実現できる．

4.3.2　知識の参照と探索

同遷移異目的ドメインにおける転移学習には，幅広い探索を行う必要があることは前項で述べた．そこで，まず，図 4.3，図 4.4 を見て欲しい．図中の S はエージェントのスタート地点を指し，G はゴール地点を表している．図 4.4 における添字付きの G は過去のタスクのゴール地点を指している．塗りつぶされている範囲が探索範囲のイメージ表している．

図 4.3 のように一般的な強化学習は，スタート地点を中心に探索範囲を均等に拡げていく．図 4.3 では，色が濃い部分がより探索され易く，色が薄い部分にいくに従いエージェントが探索しづらくなることを示している．一方，図 4.4 は [71] の手法で転移学習をした際の探索範囲をイメージしたものである．この手法では学習初期はスタート地点を中心として探索を行うが，転移学習が始まると探索の範囲が，利用した知識の経路周辺を探索していることがイメージできるだろう．このとき，図中の G_1 に向かう知識が選択されても，現在のゴールが探索範囲には含まれないため，G_1 に向かう知識が転移されることは難しい．

そこで，探索範囲を図 4.5 のように変更することを考える．このイメージでは，過去のタスクにおけるゴール周辺を探索するように設定している．こうすることで，エージェントの探索範囲が広がり，G_1 に向かう知識が転移されるべき知識として選択されるようになり，学習の高速化につながることが期待できる．しかし，利用した知識の経路周辺の探索はあきらめているため，利用した知識の経路周辺に現在のゴールがある場合は，学習が遅くなる可能性もある．これを実現するには，2 つのフェーズを必要とする．

1. 過去の知識に従うフェーズ，
2. 自由に探索するフェーズ，

である．1. は状態遷移確率が同じであるため，盲目的に過去の知識を使用することで，実現できる．2. は過去の知識に従うことなく行動を選択する必要がある．2

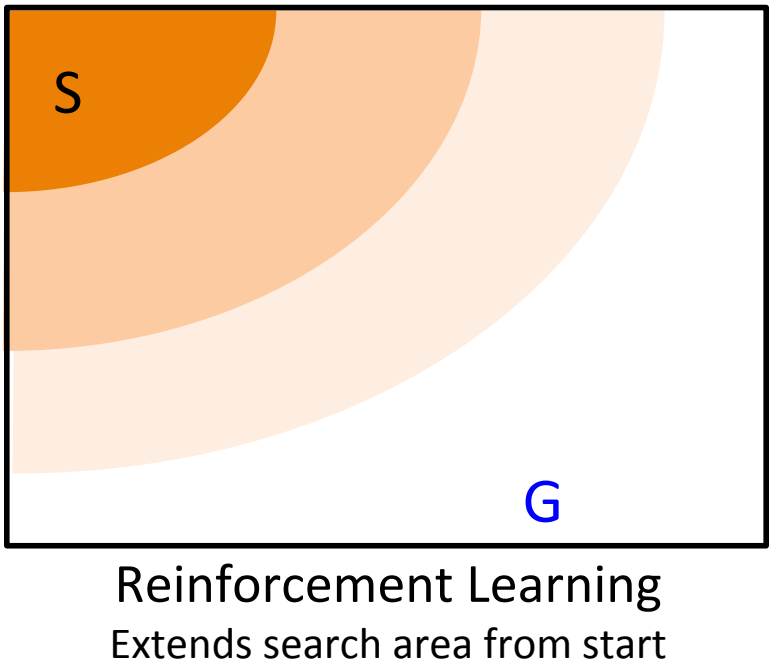


図 4.3: 強化学習による探索範囲のイメージ

つのフェーズを切り替えるこの方法を式として表すと 4.1 と表現できる．

$$\pi = \begin{cases} \pi_{\text{new}} & (t > t_0) \\ \pi_{\text{past}} & (t \leq t_0) \end{cases} \tag{4.1}$$

ここで，π_{new} は現在のタスクを学習することによって得られた行動規則，π_{past} はデータベースから選ばれた行動規則を表す．

式 4.1 はある一定時間 t_0(step) は過去の知識に従い行動を選択し，ある一定時間以後は現在の行動価値関数に従う．これは，一定時間 t_0 までに過去のタスクにおけるゴールに向かい，t_0 以降から過去のタスクにおけるゴール周辺の探索を行うことにつながる．なぜならば，現在のエージェントは過去のゴール周辺を探索する前に知識の転移が行われることが予想され，過去のゴール周辺の状態における情報をあまり収集していいないと考えられるため，途中から現在の行動価値関数に従うことで，通常の強化学習同様に過去のタスクにおけるゴール周辺を探索するようになる．

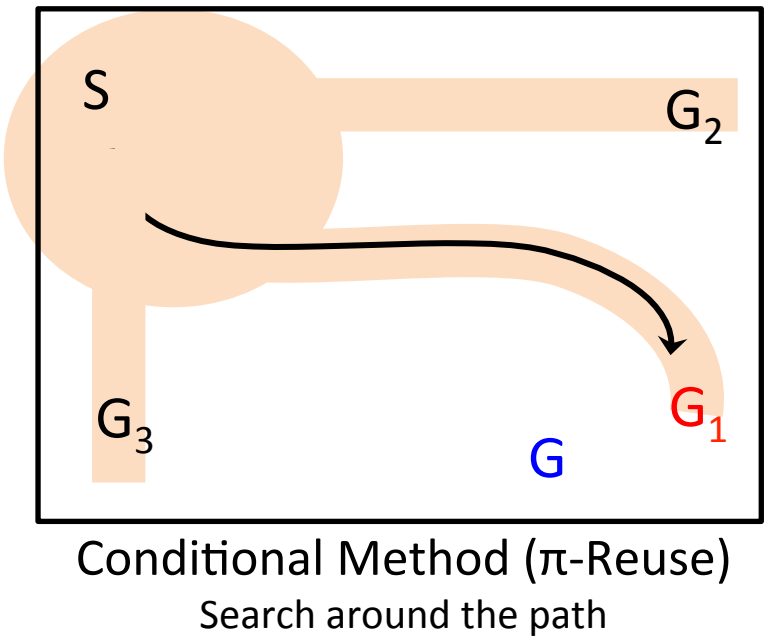


図 4.4: π-Reuse による探索範囲のイメージ

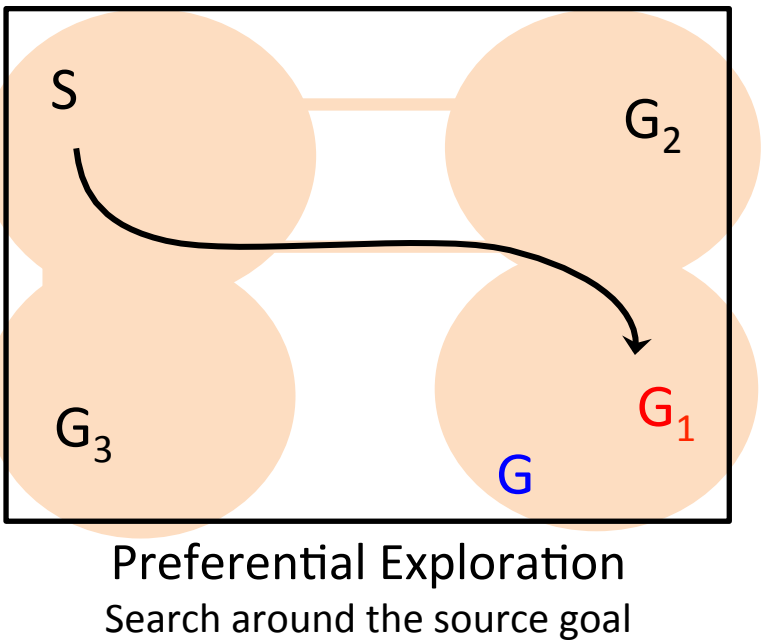


図 4.5: 提案する探索範囲のイメージ

4.4 同遷移異目的ドメインにおける転移学習を用いた強化学習のアルゴリズム

前節での検討をまとめて，アルゴリズムとして図 4.6 に示す．ここで，P は行動価値関数，V は状態価値関数を表し，F は禁止行動規則，L はデータベース，C は禁止行動規則の一致率をそれぞれ表している．また，添字 p は現在の選択していることを指し，添字 e は最も効果的であると判断された知識，添字 d はデータベースの中にある知識，♠ は，3 章における異遷移同目的ドメインに対して提案したアルゴリズムとの相違点をそれぞれ指している．また，転移させたことによってタスクを達成できた場合は禁止行動規則の一致率を δ 分だけ増加し，とそうでない場合は禁止行動規則の一致率を δ 分だけ減少させる．このように，禁止行動規則の一致率 C_{e} を変化させることにより，同遷移異目的ドメインにおいて転移すべき知識が優先的に利用されるようにアルゴリズムを改変した．

まず，禁止行動規則による知識の選択により，転移させる知識の候補が選ばれる．次に知識の参照と探索を行う手法により，一度選ばれた過去の知識を転移させてみる．知識の参照と探索を行う手法の結果を受け，転移した知識によりエージェントがタスクを達成できるならば，転移させる知識として優先順位を上げる．反対に転移させるべきではない知識と判断された場合は，転移させない知識として優先順位を下げる．これにより，多くの候補を利用し，転移すべきかすべきでないかの吟味を行うことができる．

4.5 同遷移異目的ドメインにおける転移学習による学習の高速化の検証

この節では，本章で提案した同遷移目的ドメインにおける転移学習を用いた強化学習により，現在のタスクを達成する行動規則を獲得するまでの時間が削減されるのかについて検証を行う．実験環境については，4.2 と同じであるため，ここでは説明を割愛する．実験では，禁止行動規則の一致率に傾斜配分を加えることで，選択されるべき知識が選択されるのか，また，4.6 のアルゴリズムにより学習の高速化ができるのかについて検証を行う．

//初期設定

initialize parameters P and V.

$\phi \rightarrow$ forbidden rule set F.

( ) $\rightarrow$ the latest transferred item $(P_{\mathrm{p}}, V_{\mathrm{p}}, F_{\mathrm{p}})$.

`while(` agent does not satisfy termination conditions `){`

observe state $s \in S$.

//転移させる前の強化学習

`if (` most effective item (P_e, V_e, F_e) is empty`)`

decide action $a \in A$ by following π_c.

receive reward r.

//転移後の流れ

`else`

decide action $a \in A$ by following equation 4.1. //♠

receive reward r.

//転移によりゴールできた時//♠

`if (` s is a goal for target `)`

$C_{\mathrm{e}} + \delta \rightarrow C_{\mathrm{e}}$

//転移してもゴールできなかった時//♠

`else if (` agent suspect the effectiveness `)`

$C_{\mathrm{e}} - \delta \rightarrow C_{\mathrm{e}}$

//転移させる知識の更新

`if (` a is a forbidden action `)`

add (s, a) into F.

the adjusted information $\rightarrow (P_{\mathrm{e}}, V_{\mathrm{e}}, F_{\mathrm{e}})$.

find most effective policy $(P_{\mathrm{e}}, V_{\mathrm{e}}, F_{\mathrm{e}})$ from database L

concordance rate for $(P_{\mathrm{e}}, V_{\mathrm{e}}, F_{\mathrm{e}}) \rightarrow C_{\mathrm{e}}$.

`if (` $C_{\mathrm{e}} > \theta$ `)`

$\phi \rightarrow (P_{\mathrm{e}}, V_{\mathrm{e}}, F_{\mathrm{e}})$

update P and V (actor-critic method).

}

図 4.6: 同遷移異目的ドメインにおける転移学習を用いた強化学習の流れ

表 4.2: 同遷移異目的ドメインにおける知識の選択方法により選ばれた知識

Task	Data Ω_1	データ IDΩ_2	データ IDΩ_3	データ IDΩ_4	データ IDΩ_5
Task Ω_A	347	244	26	161	165
Task Ω_B	123	105	585	143	44
Task Ω_C	0	0	0	1000	0

4.5.1 選択された知識の検証

在のタスクと過去に学習したタスクとで，状態遷移が変わらず目的状態が異なるタスクの場合，禁止行動規則は現在と過去のタスクにおいて明らかな差がつくことはない．これは，禁止行動規則が状態遷移を観測しているのと同義だからである．そこで，同遷移異目的タスクのための転移学習の提案手法が，禁止行動規則による判別が難しいものに対しても学習済みのタスクの中から，現在のタスクに必要な知識を選べるのかを検証する．

実験 2 を行い，エージェントが転移すると決断した知識について表 4.2 に示す．これは，1000[trials] 中各知識が何回転移させる知識として選ばれたかを示した表である．表中，灰色のセルは現在のタスクに適した過去のタスクを指している．

この結果では，Task Ω_A，Task Ω_B では 59%程度，Task Ω_C では 100%現在のタスクに有効な過去の知識が選択されていることが分かる．表 4.1 のように禁止行動規則のみで判断していたときは，適切なタスクを選べていなかったことをふまえると，50%以上適切な過去の知識を選択できていることから，同遷移異目的タスクおいて知識の選択が行えているといえる．

4.5.2 学習の高速化の検証

ここでは，同遷移異目的タスクにおける，行動価値関数の参照利用する一連の手法による学習の高速化について検証を行う．

実験 2 を行い，通常の強化学習による学習回数，提案手法を用いた場合の学習回数，従来研究として $\pi-$Reuse [71] による学習回数を比較する．実験の結果を表 4.3 に示す．図 4.7 は過去のタスクのゴール地点からランダム行動により探索した際のエージェントの探索範囲を示したものである．表??は各手法における学習回数を示している．灰色のセルは，通常の強化学習と比べ，T 検定によりその有意

表 4.3: 各手法における平均学習回数

Method	Task Ω_A	Task Ω_B	Task Ω_C
Original Actor-Critic	79.8	536.9	4494.4
π-Reuse	88.8	2781.3	2396.1
Proposed	100.5	259.6	83.8

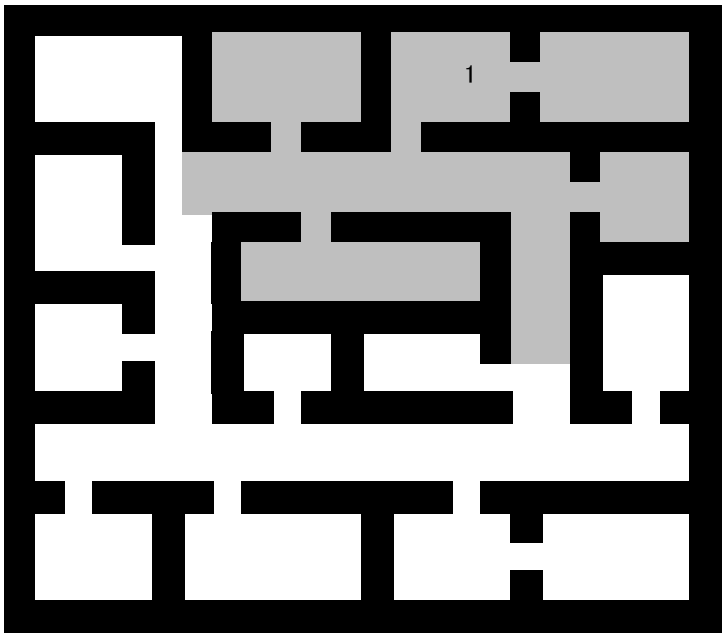
差が認められたものを表している ($p < 0.01$).

これは，各環境においてそれぞれのスタート位置（1 から 5 の数字がある地点）からランダムに行動をした際のエージェントの行動範囲を示している．灰色のセルがエージェントが行動した範囲である．

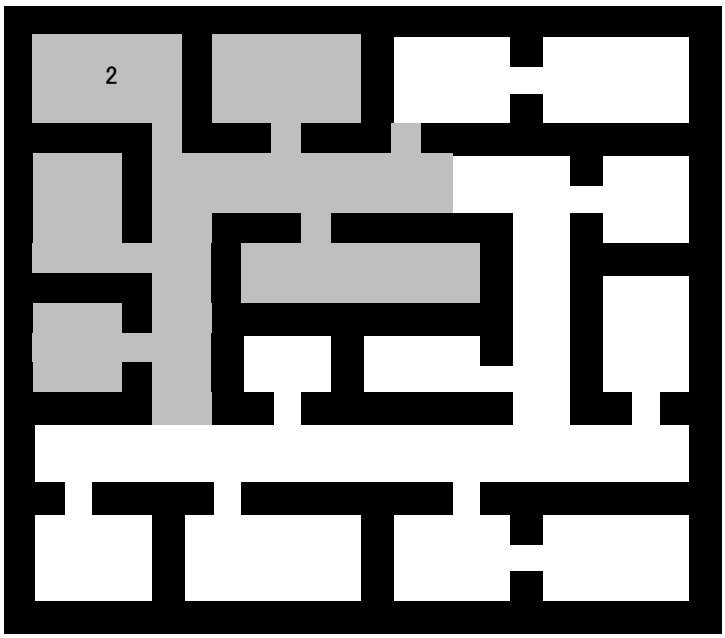
まず，$\pi-$Reuse について述べる．この手法は，Task Ω_C についてのみ学習を高速化している．これは，Task Ω_C は過去の知識においてゴールまでの経路近辺に存在するため，確率的な行動によるゴールがみつけやすかったことがいえる．そのため，Task Ω_C においては学習が高速化されたと考えられる．一方，Task Ω_B においては学習が遅くなってしまっている．これは，Task Ω_B は Task Ω_C とは異なり，経路近辺にゴールが存在しないため，確率的な行動では，Task Ω_B のゴールをみつけることができなかったといえる．

次に，提案手法について述べる．提案手法では，Task Ω_B と Task Ω_C において，学習が高速化されている．提案手法は，過去のタスクにおけるゴール周辺の探索を行いやすいため，学習が高速化されたと考えられる．このゴール周辺の探索については，図 4.7 を見てほしい．図 4.7 を見ると，Task Ω_B はデータ ID Ω_3 の知識を用いた際の行動範囲に含まれ，TaskΩ_C はデータ ID Ω_4 の知識を用いた際の行動範囲にそれぞれ含まれていることが分かる．特に，TaskΩ_C については，かなりの学習回数が削減できている．このことはおそらく，過去の知識を一定ステップだけ利用したことで，図 4.7 で示されているランダム行動の探索範囲に現在のゴールが存在したため，学習回数が削減されたと推測される．また，π-Reuse は，経路の近くにゴールが存在したため，学習が早くなったと考えられる．しかし，確率的に過去の知識と現在の知識とを入れ替えているため，運良く過去の経路からはずれ，ゴールに到達することが少なかったことが考えられる．その分，提案法の方が学習回数を削減する結果となったと考えられる．このことから，提案手法により過去のタスクにおけるゴール周辺の探索が行われやすくなった結果，現在のタスクの学習が高速化されたと考えられる．

最後に Task Ω_A についてだが，これは，図 4.7 を見てもらうと分かるのだが，ランダム行動により既にゴール地点をみつけることが可能であることがわかる．この場合，転移学習を使わなくても速い段階でエージェントは学習を完了させることが可能であるといえる．そのため，提案手法や $\pi-$Reuse を用いた手法による学習の高速化の効果は期待できないといえる．

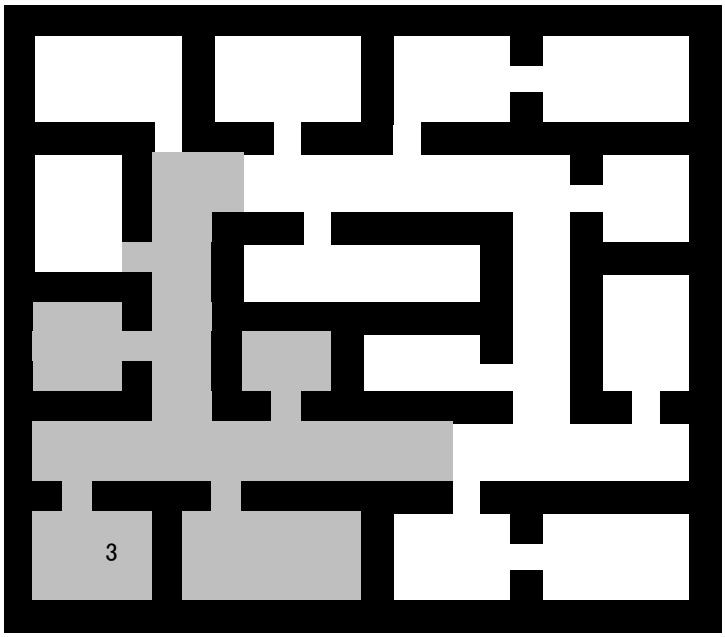


(1) データ ID Ω_1(Task Ω_A にゴールが近い)

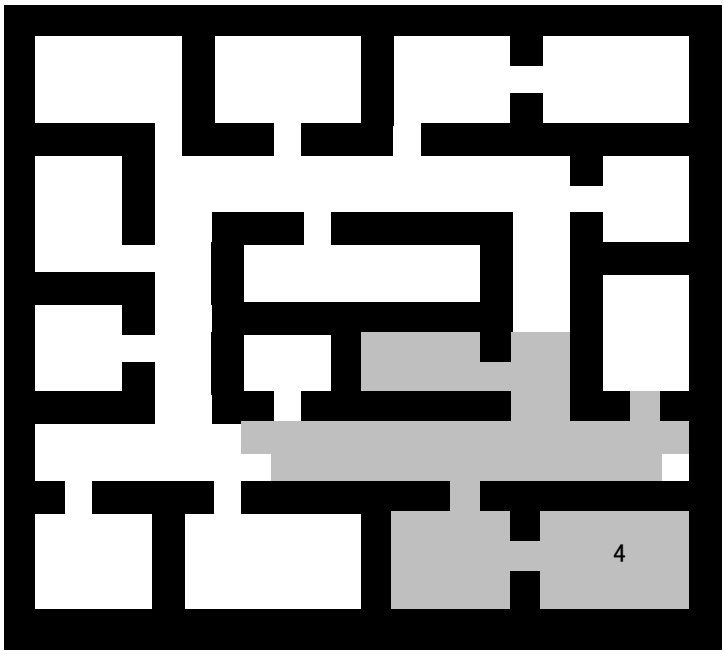


(2) データ ID Ω_2(Task Ω_A にゴールが近い)

図 4.7: 過去のタスクのゴールからのランダム行動による探索範囲

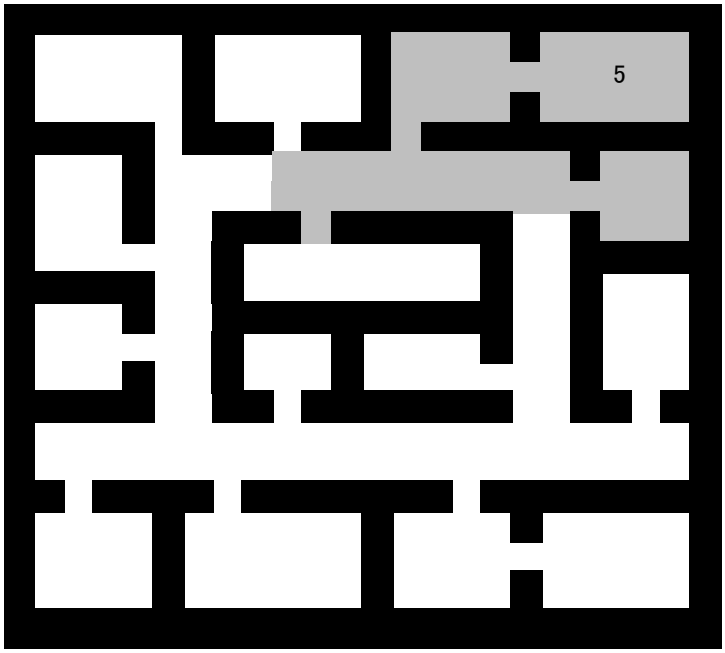


(3) データ ID Ω_3(Task Ω_B にゴールが近い)



(4) データ ID Ω_4(Task Ω_C にゴールが近い)

図 4.7: 過去のタスクのゴールからのランダム行動による探索範囲 (続き)



(5) データ ID Ω_5 どのゴールからも離れている

図 4.7: 過去のタスクのゴールからのランダム行動による探索範囲 (続き)

4.6 結び

本章では，同遷移異目的ドメインにおける転移学習を用いた強化学習について検討 を行った．まず，異遷移同目的ドメインにおける知識の選択手法を実験により使用してみたところ，このドメインでは，状態遷移確率が同じであるため，過去のタスク間において明確な差異をつけることはできないことを確認した．次に，知識を一度使ってみるという [71] を参考にして，禁止行動規則を用いた知識の選択方法の改善について検討を行った．また，同遷移異目的ドメインにおいては，幅広い状態を探索する必要があるため，転移学習として 1. 過去の行動価値関数に従う，2. 現在の行動価値関数に従い自由に探索をする，という 2 つのフェーズからなる知識の参照と探索を行う手法を提案した．データベースから同遷移異目的ドメインにおける知識の選択ができるようになったことを確認し，さらに本章の提案により学習の高速化につながることを実験を通して確認された．

第5章　総論

5.1　本研究で得られた成果

強化学習は，与えられたタスクにおいて環境とのインタラクションを通じて自律的にタスクを達成する行動規則を獲得する枠組みである．しかし，自律的な行動規則獲得のためには，多くの学習時間が必要となることが問題とされている．この問題を解決するため，エージェントがこれまでの経験により獲得した知識を現在の学習に利用する転移学習が研究されている．強化学習において転移学習を適用するためには，過去のエージェントと現在のエージェントとの関係，過去のタスクと現在のタスクの関係について考慮しなければ，学習を高速化につながる転移学習とはならない．

本研究では，転移学習を用いることにより強化学習の学習を高速化を図るための検討を行った．具体的には，過去のタスクと現在のタスクの関係を，異遷移同目的タスクと同遷移異目的タスクの 2 つにタスクの関係を分割した．異遷移同目的タスクについては，禁止行動規則を用いて転移させる知識の選択を行い，選択された過去の知識と現在の知識の条件付き加重平均をとることにより，学習の高速化が図れることを示した．

一方，同遷移異目的タスクについては，異遷移同目的タスクのように知識の選択ができないため，禁止行動規則に加え，過去の知識を参照した結果もふまえることで，現在のタスクに利用できる知識の選択手法を提案した．また，過去の知識を一定時間参照することにより，従来手法の $\pi-$Reuse よりも広い探索範囲をエージェントが探索できることを確認した．そして，この手法による学習の高速化についても実験により示した．

これらの手法は，検討を簡単にするため，同一エージェント間における転移学習とした．

5.2　今後の課題

今後の課題としては，以下に示すことが挙げられる．

- 同一エージェントだが，環境もタスクも異なる場合の転移学習はどうなるか，
- 類似エージェント，マルチエージェントにおける転移学習，
- 現実のモデルでも利用できるのか (会話ロボットなど)，

といった，問題が残されている．

環境もタスクも異なる場合というのは，今回検討した 2 つのドメインの複合したドメインにおける転移学習のことを指している．これは，エージェントが自動車の自律走行を行う際，地図を知識としてエージェントが保持している場合，スタート地点や目的地によって地図の使い方が変わってくるような状況などが考えられる．こうした状況にも対処できるようにすることで，実現の可能性が高まってくると考えられる．

強化学習における転移学習はエージェントや環境，タスクの性質によって，その手法が大きく異なることが本研究からも分かる．そのため，人間社会で起こりうることをできる限り網羅していないと，転移学習を用いた強化学習を有用なものとして実現させることは困難であると考えられる．特に，人間の生活環境は常に変化し続けるため，動的環境への対応は必須である．また，前述した課題をシミュレーションで解決した後にも，シミュレーション環境と現実環境とのギャップが存在するため，さらなる課題が発見されるであろう．

// 実社会で利用されているロボットの例

[1] ロボット家電, http://www.sharp.co.jp/cocorobo/communication/, Sharp 株式会社. (2012 年 12 月 18 日最終アクセス).

[2] ダ ヴィン チ 導 入 に あ たって, http://www.matsunami-hsp.or.jp/iryou_jyohou/davinci/index.html, 社会医療法人 蘇西厚生会 松波総合病院. (2012 年 12 月 18 日最終アクセス).

[3] 介 護 ロ ボット 紹 介, http://www.kaigo-robot-kanafuku.jp/category/1438992.html, 公益社団法人神奈川福祉サービス振興会, (2012 年 12 月 18 日最終アクセス).

// 強化学習に関する文献

[4] Richard S. Sutton, Andrew G. Barto 著, 三上貞芳, 皆川雅章 訳; 強化学習, 森北出版, 2000.

[5] Leslie Pack Kaelbling, Michael L. Littman and Andrew W. Moore; Reinforcement Learning — A Survey, Journal of Artificial Intelligence Research, vol.4, pp.237–285, 1996.

[6] Witten I. H.; An Adaptive Optimal Controller for Discrete-Time Markov Environments, Information and Control, Vol.34, pp. 286–295, 1977.

[7] A. G. Barto, R. S. Sutton, and C. W. Anderson; Neuronlike elements that can solve difficult learning control problems, IEEE Trans. on Systems, Man, and Cybernetics, Vol.13, pp. 835–846, 1983.

[8] 木村元, 小林重信; 確率的 2 分木の行動選択を用いた Actor-Critic アルゴリズム, 計測自動制御学会論文集, Vol.33, No.1, pp.1147–1155 ,1997.

[9] 木村元; 適性度の履歴を用いた自然勾配 Actor-Critic 法, 計測自動制御学会第 19 回自律分散システムシンポジウム, pp.67–72, 2007.

[10] Vijay Konda, John Tsitsiklis; On Actor-Critic Algorithms, SIAM Journal on Control and Optimization, Vol.42, No.4, pp1143–1166, 2003.

[11] 森健, 吉本潤一郎, 石井信; 確率的方策勾配法に基づく actor-critic 法と連続システムの制御への応用, 電子情報通信学会技術研究報告, NC2002-226, pp.137–142, 2003.

[12] 加藤新吾, 松尾啓志; 動的環境下における Profit Sharing, 電子情報通信学会論文誌, D-I, Vol.J84-D-I, No.7, pp.1067–1075, 2001.

[13] 中村泰, 石井信; 自然方策勾配法に基づくオフポリシー型強化学習, 電子情報通信学会技術研究報告, NC2004-191, pp.131–136, 2004.

[14] 畝見達夫; 強化学習, 人工知能学会誌, Vol.9, No.6, pp.830–836, 1994.

// マルコフ決定過程に関する文献

[15] 涌田和芳; 不完全状態観測のセミマルコフ決定過程. 日本オペレーションズ・リサーチ学会論文誌, Vol.24, No.5, pp.95–109, 1981.

[16] Lovejoy, W. S.; A Survey of Algorithmic Methods for Partially Observed Markov Decision Processes. Annuals of Operations Research 28, pp.47–65, 1991

[17] 木村元, Leslie Pack kaelbling; 部分観測マルコフ決定過程下での強化学習. 人工知能学会誌, Vol.12, No.6, pp.822–830,1997.

// 状態空間の構成方法に関する研究

[18] 宮本行庸, 上原邦昭; 特徴構成法を用いた Q 学習の効率改善. 情報処理学会論文誌 数理モデル化と応用 40(SIG9(TOM2)), pp.62–71, 1999.

[19] 岩崎秀樹, 末田直道; 強化学習における自己組織化マップを用いた状態空間の自律的構成法. 第 19 回人工知能学会全国大会講演論文集, 1D3-05, 2005.

[20] 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次的構成. 日本ロボット学会誌, Vol.17, No.1, pp.118–124, 1999.

[21] Minoru Asada, Shoichi Noda, and Koh Hosoda; Action-Based Sensor Space Categorization for Robot Learning. In Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems '96 (IROS 96), Vol.3, pp.1502–1509, 1996.

// パラメータ調整による学習の高速化に関する研究

[22] 亀井圭史, 石川眞澄; 移動ロボットの強化学習パラメータの環境依存性, 電子情報通信学会技術研究報告. NC, 102 巻 628 号, Vol.105, No.659, pp.61–66, 2006.

[23] 溝上裕之, 小林邦和, 呉本尭, 大林正直; TD 誤差に基づく教科学習のめたパラメータ学習法. 電気学会論文誌 C, 129 巻 9 号, pp.1730–1736, 2009.

[24] 高橋哲也, 安達雅春; アクター・クリティック型強化学習における学習率の非線形スケジューリング. 電子情報通信学会技術研究報告, NLP2003-178, pp.13–18, 2004.

[25] 吉田和子, 石田信; 強化学習における exploration と exploitation の制御. 電子情報通信学会技術研究報告, NC2001-28, pp.41–47, 2001.

[26] Komla A. Folly and Ganesh K. Venayagamoorthy; Effects of Learning Rate on the Performance of the Population Based Incremental Learning Algorithm. In Proceedings of International Joint Conference on Neural Networks, pp.861–868, 2009.

[27] Masayuki Hara, Masashi Inoue, Jian Huang and Tetsuro Yabuta; Study on Motion Forms of Mobile Robots Generated by Q-Learning Process Based on Reward Databases. IEEE International Conference on Systems, Man, and Cybernetics 2006, pp.5112–5117, 2006.

[28] 宮崎和光, 山村雅幸, 小林重信; 強化学習における報酬割当ての理論的考察. 人工知能学会誌, Vol9, No.4, pp.580–587, 1994.

[29] 尾川 順子, 並木 明夫, 石川 正俊; 学習進度を反映した割引率の調整. 電子情報通信学会技術研究報告. NC, pp.73–78, 2003.

// 柔軟な振る舞いを行うロボットの動作獲得に関する研究

[30] 岡本充義, 山口智洋, 谷内田正彦; 多戦略学習手法 MS-RL: 環境変動下におけるロバストな学習エージェントの実現. 電子情報通信学会技術研究報告, AI98-73, pp.31–38, 1999.

[31] 森本淳, 銅谷賢治; ロバスト強化学習. 電子情報通信学会技術研究報告, NC2000-49, pp.59–66, 2000.

[32] 時田陽一, 中村泰, 吉本潤一郎, 石井信; モデル誤差を考慮した強化学習手法による実ロボットの適応制御. 電子情報通信学会技術研究報告, NC2005-154, pp.19–23, 2006.

[33] 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, Vol.18, No.5, pp.706–712, 2000.

// 学習の高速化に関する研究

[34] 田淵一真, 谷口忠大, 椹木哲夫; 模倣学習と強化学習の調和による効率的行動獲得. 第 20 回人工知能学会全国大会, 3C1-2, 2006.

[35] 大西弘将, 横井博一; 二足歩行ロボットのための仮想学習システム. 電子情報通信学会技術研究報告, NC2006-57, pp.7–11, 2006.

[36] 森紘一郎, 山名早人; 強化学習並列化による学習の高速化. 電子情報通信学会技術研究報告, AI2003-91, pp.59–64, 2004.

[37] 大東優, 大森隆司, 森川幸治, 岡夏樹; 予測ベース強化学習に基づくゲーム学習の加速–プランニング行動の発生に向けて–. 電子情報通信学会技術研究報告, NC2002-113, pp.61–66, 2003.

[38] 鮫島和行, 片桐憲一, 銅谷賢治, 川人光男; 複数の予測モデルを用いた強化学習による非線形制御. 電子情報通信学会論文誌, Vol.J84-D-II, No.9, pp.2092-2106, 2001.

// 過去の事例を活用する学習手法に関する研究

[39] 森中雄, 大原剛三, 馬場口登, 北橋忠宏; 事例間の類似性に基づく負例の生成によるデータベースからの知識獲得. 人工知能学会誌, Vol.15, No.5, pp.862–869, 2000.

[40] 小幡琢磨, 佐々木洋輔, 久保村千明, 亀田弘之; 事例を経験として蓄積し利用する強化学習手法の提案. 電子情報通信学会技術研究報告, TL2005-97, PRMU2005-232, 2006.

[41] 山口明彦, 杉本徳和, 川人光男; 回避行動の再利用メカニズムを備えた強化学習手法と多関節ロボットの全身運動学習への応用. 日本ロボット学会誌, Vol.27, No.2, pp.209–220, 2009.

[42] 宮崎和光, 坪井創吾, 小林重信; 罰を回避する合理的政策の学習. 人工知能学会誌, Vol.16, No.2, pp.185–192, 2002.

[43] 石川浩一郎, 櫻井彰人, 藤波努, 國藤進; 複数の状態行動価値表を用いた R 学習の高速化. 電気学会論文誌 C, 電子 · 情報 · システム部門誌, 126(1), pp.72–82, 2006.

// 強化学習の問題点を指摘している文献

[44] Jeffery A. Clouse and Paul E. Utgoff; A Teaching Method for Reinforcement Learning. In Proceedings of the Ninth International Workshop on Machine Learning, pp.92-110, 1992

[45] 荒井幸代; マルチエージェント強化学習: 実用化に向けての課題 · 理論 · 諸技術との融合. 人工知能学会誌, Vol.16, No.4, pp.476-481, 2001.

// 同一エージェント間における転移学習に関する研究

[46] David Andre and Stuart J. Russell; State Abstraction for programable reinforcement learning agents. In Proceedings of the 18th National Conference on Artificial Intelligence, pp.119–125, 2002.

[47] Minoru Asada, Shoichi Noda, Sukoya Tawaratsumida, and Koh Hosoda; Vision-bsed behavior acquisition for a shooting robot by using a reinforcement learning. In Proceedings of IAPR/IEEE Workshop on Visual Behaviors-1994, pp.112–118, 1994.

[48] Mehran Asadi Manfred Huber; Effective control knowledge transfer through learning skill and representation hierarchies. In Proceedings of the 20th International Joint Conference on Artificial Intelligence, pp.2054–2059, 2007.

[49] Christopher G. Atkeson and Juan C. Santamaria; A comparison of direct and model-based reinforcement learning. In Proceedings of the 1997 International Conference on Robotics and Automation, pp.3557–3564, 1997.

[50] Kimberly Gerguson and Sridhar Mahadevan; Proto-transfer learning in Markov decision processes using spectral method. In Proceedings of the ICML-06 Workshop on Structural Knowledge Transfer for Machine Learning, 2006.

[51] Alessandro Lazaric; Knowledge Transfer in Reinforcement Learning. PhD thesis, Politecnico di Milano, 2008.

[52] Michael G. Madden and Tom Howley; Transfer of experience between reinforcement learning environments with progressive difficulty. Artificial Intelligence Review, 23(3–4), pp.375–398, 2004.

[53] Balaraman Ravindran and Andrew G. Barto; Relativized options: choosing the right transformation. In Proceedings of the 20th International Conference on Machine Learning (ICML2003), pp.608–615, 2003.

[54] Oliver G. Selfridge, Richard S. Suttton, and Andrew G. Barto; Training and tracking in robotics. In Proceedings of the 9th International Joint Conference on Artificial Intelligence, pp.670–672, 1985.

[55] Alexander A. Sherstov and Peter Stone; Improving action selection in MDP'S via knowledge transfer. In Proceedings of the 20th National Conference on Artificial Intelligence, pp.1024–1029, 2005.

[56] Satinder P. Singh; Transfer of learning by composing solutions of elemental sequential tasks. Machine Learning, 8:323–339, 1992.

// 類似エージェント間における転移学習に関する研究

[57] Bikramjit Banerjee and Peter Stone; General game learning using knowledge transfer. In The 20th International Joint Conference on Artificial Intelligence, pp.672–677, 2007.

[58] Tom Croonenborghs, Kurt Driessens, and Maurice Bruynooghe; Learning relational options for inductive transfer in relational reinforcement learning. In Proceedings of the 17th Conference on Inductive Logic Programming, pp.88–97, 2007.

[59] Carlos Guestrin, Daphne Koller, Chris Gearhart, and Neal Kanodia; Generalizing plans to new environments in relational MDPs. In International Joint Conference on Artificial Intelligence (IJCAI-03), pp.1003–1010, 2003.

[60] George Konidaris and Andrew Barto: Autonomous shaping; knowledge transfer in reinforcement learning. In Proceedings of the 23rd International Conference on Machine Learning, pp.489–496, 2006.

[61] George Konidaris and Andrew G. Barto; Building portable options: skill transfer in reinforcement learning. In Proceedings of the 20th International Joint Conference on Artificial Intelligence, pp.895–900, 2007.

[62] Jan Ramon, Kurt Driessens, and Tom Croonenborghs; Transfer learning in reinforcement learning problems through partial policy recycling. In Proceedings of the 18th European Conference on Machine Learning, pp.699–707, 2007.

[63] Manu Sharma, Michael Holmes, Juan Santamaria, Arya Irani, Charles Isbell, and Ashwin Ram; Transfer learning in real-time strategy games using hybrid CBR/RL. In Proceedings of the 20th International Joint Conference on Artificial Intelligence, pp.1041–1046, 2007.

[64] Matthe E. Taylor and Peter Stone; Cross-domain transfer for reinforcement learning. In Proceedings of the 24th International Conference on Machine Learning, pp.879–886, 2007.

[65] Matther E. Taylor, Peter Stone, and Yaxin Liu; Transfer learning via intertask mappings for temporal difference learning. Journal of Machine Learning Research, 8(1):2156–2167, 2007.

[66] Matthew E. Taylor, Shimon Whiteson, and Peter Stone; Transfer via intertask mappings in policy search reinforcement learning. In The 6th International Joint Conference on Autonomous Agents and Multiagent Systems, Article No.37, 2007.

[67] Matthew E. Taylor, Gregory Kuhlmann, and Peter Stone; Autonomous Transfer for reinforcement Learning. In Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems, Vol.1, pp283–290, 2008.

[68] Lisa Torrey, Trevor Walker, Jude W. Shavlik, and Richard Maclin; Using advice to transfer knowledge acquired in one reinforcement learning task to

another. In Proceedings of the 16th European Conference on Machine Learning, pp.412–424, 2005.

[69] Lisa Torrey, Jude W. Shavlik, Trevor Walker, and Richard Maclin; Skill acquisition via transfer learning and advice taking. In Proceedings of the 17th European Conference on Machine Learning, pp.425–436, 2006.

[70] Lisa Torrey, Jude W. Shavlik, Trevor Walker, and Richard Maclin; Relational macros for transfer in reinforcement learning. In Proceedings of the 17th Conference on Inductive Logic Programming (ILP 2007), pp.254-268, 2007.

// マルチタスクに関する研究

[71] Fernando Fernández and Manuela Veloso; Probabilistic policy reuse in a reinforcement learning agent. In Proceedings of the 5th International Conference on Autonomous Agents and Multiagent Systems, pp.720–727, 2006.

[72] David Foster and Peter Dayan; Structure in the space of value functions. Machine Learning, 49, Issue 2–3, pp.325–346, 2002.

[73] Neville Mehta, Sriraam Natarajan, Prasad Tadepalli, and Alan Fern; Transfer in variable-reward hierarchical reinforcement learning. MachineLearning, 73(3), pp.289–312, 2008.

[74] Theodore J. Perkins and Doina Precup; Using options for knowledge transfer in reinforcement learning. Technical Report UM-CS-1999-034, The University of Massachusetts at Amherst, 1999.

[75] Funlade T. Sunmola and Jeremy L. Wyatt; Model transfer for Markov decision tasks via parameter matching. In Proceedings of the 25th Workshop of the UK Planning and Scheduling Special Interest Group(PlanSIG 2006), 2006.

[76] Aaron Wilson, Alan Fern, Soumya Ray, and Prasad Tadepalli; Multitask reinforcement learning: a hierarchical Bayesian approach. In Proceedings of the 24th International Conference on Machine Learning (ICML07), pp.1015–1022, 2007.

[77] Thomas J. Walsh, Lihong Li, and Mchael L. Littman; Transferring state abstractions between MDPs. In ICML Workshop on Strucrtural Knowledge Transfer for Machine Learning, 2006.

[78] Fumihide Tanaka and Masayuki Yamamura; Multitask reinforcement learning on the distribution of MDPs. 2003 IEEE International Symposium on Computational Intelligence in Robotics and Automation, Vol.3, pp.1108–1113, 2003.

// 転移学習についてサーベイを行っている文献

[79] Matthew E. Taylor and Peter Stone; Transfer Learning for Reinforcement Learning Domains: A Survey. The Journal of Machine Learning Research, Vol.10, pp.1633–1685, 2009.

[80] S. J. Pan and Q. Yang.; A survey on transfer learning. Technical Report HKUST-CS08-08, Dept. of Computer Science and Engineering, Hong Kong Univ. of Science and Technology, 2008.

[81] Emilio Soria Olivas, JoseéDavid Martín Guerrero, Marcelino Martinez-Sober, Jose Rafael Magdalena- Benedito, and Antonio José Serrano López; Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques. IGI Global, pp.242–264, 2009.

[82] 神嶌敏弘; 転移学習. 人工知能学会誌, Vol.25, No. 4, pp.572-580, 2010.

// マルチエージェントに関する文献

[83] 高玉圭樹: マルチエージェント学習 -相互作用のなぞに迫る-. コロナ社, 2003.

[84] 生天目章; マルチエージェントと複雑系. 森北出版, 1998.

[85] 上田完次 編著; 創発とマルチエージェントシステム. 培風館, 2007.

[86] 石原聖司, 五十嵐治一; マルチエージェント系における方策勾配法—追跡問題—. 電子情報通信学会技術研究報告, AI2002-58, pp.65–70, 2003.

[87] 神尾正太郎, 伊庭斉志; マルチエージェント協調作業のためのランダムサンプリングを用いた経路プランニングアルゴリズム. 電子情報通信学会論文誌 D, Vol.J89-D, No.2, pp.250–260, 2006.

// 教育心理学における転移学習に関する文献

[88] Thorndike, E. L and Woodworth, R. S.; The influence of improvement in one mental function upon the efficiency of other function. Psychological Review, 8, pp.247–261, 1901.

[89] 三枝浩, 石川智博, 加藤幸一; 技術科教育における問題解決能力の育成について. 群馬大学教育実践研究別刷, 第 27 号, pp.145–162, 2010.

[90] 権裕善, 藤村宣之; 同年齢児童の共同はいつ有効であるのか: 比例的推理の方略レベルが異なるペアの相互作用. 教育心理学研究, 52(2), pp.148–158, 2004.

[91] 寺尾敦; 数学的問題解決における転移を促進する知識の獲得について. 教育心理学研究, 46(4), pp.461–472, 1998.

[92] 佐藤正二, 佐藤容子; 幼児における自己教示訓練と方略転移. 日本教育心理学会総会発表論文集, (24), pp. 772–773.

[93] 植友理; 学習方略は教科間でいかに転移するか: 「教訓帰納」の自発的な利用を促す事例研究から. 教育心理学研究, (58), pp.80–94, 2010.

[94] Gerald J. Calais; Haskell's Taxonomies Of Transfer Of Learning: Implications For Classroom Instruction. National Forum of Applied Educational Research Journal, Vol.20, No.3, pp.1–8, 2006.

// 転移学習や強化学習の高速化に関するその他の文献

[95] Hoshino Y., Kamei K.; A Proposal of Reinforcement Learning System to Use Knowledge Effectively. SICE Annual Conference 2003, Vol.2, pp.1582-1585, 2003.

[96] S. Thrun.; Is learning the $n-$th thing any easier than learning the first?. In Advances in Neural Information Processing Systems 8, pp.640-646, 1996.

[97] 松井藤五郎, 犬塚信博, 世木博久, 伊藤英則; 強化学習結果の再構築への概念学習の適用. 人工知能学会論文誌, Vol.17, No.2, pp.135-144, 2002.

[98] 大村英史, 片上大輔, 新田克己; 学習結果のカテゴリ化によるマルチタスクへの適応. 第 20 回人工知能学会全国大会, 1B3-4, 2006.

[99] 大東優, 大森隆司, 石川悟, 森川幸治; 部分環境モデルの組み合わせによるすばやいタスクモデルの構築と多重タスク学習への適用. 電子情報通信学会技術研究報告, NC2004-219, 2005.

[100] 市瀬龍太郎, 武田英明, 本位田真一; 階層的知識間の調整規則の学習. 人工知能学会論文誌, Vol.17, No.3, pp.230–238, 2002.

謝辞

本論文は，著者が三重大学大学院工学研究科博士後期課程時に行った研究をまとめたものである．本論文を進めるにあたり，懇切丁寧な御指導と御督励を賜った三重大学の林照峯名誉教授，鶴岡信治教授，北英彦准教授，高瀬治彦准教授，川中助教に感謝いたします．さらに，筆者の研究に関して，いろいろご教示いただいた三重大学工学部電気電子工学科小林英雄教授，同大学工学部情報工学科木村文隆教授，同大学工学部電気電子工学科森香津夫教授に深く感謝いたします．また，日頃熱心に討論していただいた計算機工学研究室，情報処理研究室の皆様方に厚く御礼申しあげます．

最後に，本論文をまとめるにあたり，助言，討論，その他お世話になったすべての方々に感謝いたします．