変換不変性を用いた経験の抽象化と内発的動機づけに基づく強化学習 *

増山 岳人 *1, 山下 淳 *2, 淺間 一 *2

Reinforcement Learning Based on Intrinsic Motivation and Temporal Abstraction via Transformation Invariance

Gakuto MASUYAMA*1, Atsushi YAMASHITA and Hajime ASAMA

*1 Department of Precision Engineering, Faculty of Engineering, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

Bottom-up processes have received much attention in unsupervised and developmental learning research domain. In contrast, effectiveness of top-down deeming on acquisition of adaptive behavior is discussed in this paper. Successful experience in the past, or a skill that could be expected to be reused successfully in a novel environment is stored in memory. Then abstract environment recognition via geometric transformation invariance is introduced to measure the reproducibility of executed skill in a novel environment. Additionally, reproducibility of skill in the environment is utilized to make up intrinsic motivation that drives the agent to active conceptualization of search space. It enables the agent to relativize current skill execution robustly in diverse environments. Useful characteristics of top-down deeming process are implemented on reinforcement learning and discussed through simulation experiments in grid world. The results demonstrate acceleration of learning progress by active conceptualization of environment. Additionally, it is shown by experiments for scaled environment that subjective anticipation could bring in consistent strategy of exploration and exploitation. Eligibility trace is also introduced for skill utility problem and it is shown that the traces regarding actions and skills could preserve learning performance for diverse skill settings.

Key Words : Reinforcement Learning, Learning Control, Knowledge Engineering, Robot

1. 緒　　言

十分な事前情報を用いることなく，環境との相互作用の過程において漸増的に学習する自律的なロボットの設計は，社会的な場面へのロボットの導入のための重要な研究課題である．タスク，環境，ロボットのダイナミクス等に関する明確な事前情報を要求せず，試行錯誤的に学習を行うことができる強化学習[(1)]はこのような問題に対して親和性が高く，これまで様々な研究が行われてきた．近年では発達ロボティクス[(2)]の文脈において，強化学習への内発的動機づけの導入が注目を集めている．動物は未知のタスクや環境に対して迅速な意思決定を行い，適応的に振る舞うことができる．このような機能が実現される大きな要因は，学習者を包括的な知識の獲得に向け駆動する内発的動機づけのシステムであると言われている．心理学研究においては動機づけは外発的動機付けと内発的動機づけに分けられ，従来の強化学習で扱われてきたのは外発的報酬であり，学習者を目的的な行動に駆動する．他方，内発的動機づけは好奇心や興味のような，行動自体が目的となるような行動をもたらす．

内発的に動機づけられた行動は特定の一つのタスクのためではなく，幅広い場面で利用できる能力の獲得にあるとの考えに基づいて，強化学習手法に内発的動機づけを組み込む研究が行われてきた．Barto らの提案した Intrinsically Motivated Reinforcement Learning（IMRL）[(3)(4)]では，option による状態及び行動の時間的な抽象化と，有用なスキ

* 原稿受付　2012 年 8 月 27 日
*1 東京大学大学院工学系研究科精密工学専攻 （〒 113-8656 東京都文京区本郷 7-3-1）
*2 正員, 東京大学大学院工学系研究科精密工学専攻 （〒 113-8656 東京都文京区本郷 7-3-1）

Email: masuyama@robot.t.u-tokyo.ac.jp

ル獲得のための salient event に対する内発的報酬が導入されている．行動の時間的拡張である option[5]を獲得することで，外発的報酬のみを用いる場合より効率的な学習が可能であることが示されている．また，Oudeyer らは確率的に記述することが困難な環境の特性に起因する学習の停滞に対して，Intelligent Adaptive Curiosity（IAC）と呼ばれる動機づけシステムを提案した[6]．IAC は学習の進行を観測し，それを最大化することでエージェントを未学習かつ予測が可能な領域へ駆動する機能を実現した．特に IMRL と同様，option の枠組みにおける行動の階層化は活発に研究されており，タスクと独立した内発的動機づけを利用することで学習の促進が可能であることが示されてきた[7][8]．

これまでの内発的に動機づけられた強化学習に関する研究は，主に有用なスキルの獲得や学習の高速化を目的とした，センサデータ，その特徴量，そしてそれらの分析に基づくボトムアップ処理に着目している．しかしながら，発達ロボティクスの明確な目標の一つである人間の情報処理過程においてはボトムアップ処理とトップダウン処理が併用されることが知られている[9]．ボトムアップ処理はデータ駆動型で，末梢系からの連続な入力とその分析結果から得られる特徴量に基づく受動的かつ逐次的な処理である．他方，トップダウン処理は概念駆動型であり，大脳中枢の抽象的な知識が何らかの概念を表す離散的な記号に基づく，能動的かつ探索的な処理である．知識とボトムアップ処理により得られる解析結果との対応は必ずしも明確ではない．そこで，トップダウン処理は知識に基づく “期待”を作り出し，さらにその期待を実際に確認するためのバイアスを情報処理システムに加える．つまり，トップダウン処理は環境に対する積極的な概念化を促す機能をもつといえる．人間の情報処理過程におけるトップダウンなバイアスは物体認識の促進と関連付けられており[10]，また，ある種の認知バイアスは動機づけと関連すると言われている[11]．本論文では従来の内発的に動機づけられた強化学習の枠組みでは注目されてこなかった，内発的な動機づけと関連付けられる上述のトップダウン処理の特異的な性質に着目している．

発達の初期段階における有用なスキルの獲得を目的とするボトムアップ処理は，自律ロボット設計の方法論を議論する上で極めて重要なものである．しかしながら，スキルが蓄積され行動に関する階層的な構造が拡張されるにしたがって，時間・空間的な探索空間の急激な拡大が起こる．その結果，エージェントの発達過程が進むにつれて見かけ上，上位層における学習速度の低下という問題が起こる．本論文では，この問題に対してトップダウン処理を強化学習へ実装し，過去の成功経験に関する知識に基づく期待によって探索戦略にバイアスを与え学習を加速する手法を提案する．新たな環境における過去の経験の再現性に対する期待によって探索空間を狭め，期待に基づく概念化を積極的に試みることで未知環境への適応が促進される．以後，過去に経験したある有限の行動系列と，それに伴って観測されたセンサ情報に関する抽象的知識の組をスキルと呼ぶこととする．ここでの抽象的知識とはスキルの実行に伴って観測されるセンサ情報の時系列を低次元化した量である．それぞれのスキルに対応する抽象的知識は，新たな環境におけるスキルの実行結果の過去の経験に対する再現性を評価するために用いられる．未知環境においてスキルは試行錯誤によって選択され，過去の成功経験に対する高い再現性をもつスキルが内発的報酬によって強化される．そして行動選択過程において，内発的に動機づけられたスキルの実行がその状態において適切である（あるいは適切でない）という期待に基づくバイアスが加えられる．内発的な動機づけはタスクと独立しており，それによって強化されたスキルの実行が新たな環境において正の外発的報酬を得る保証を与えるものではない．しかし，成功経験から抽出されたスキルは，行動を整合性のある形式で時間的に拡張する．スキルの実行が高い経験の再現性，正の報酬の獲得，あるいは高い価値をもつ状態への遷移等によって強化されたとき，スキルは行動価値上に埋め込まれ，探索の中心となる経路が構成される．その結果，知識によって探索空間は縮減し，指向性のある行動選択により学習の高速化が図られる．また，以上のアイディアを実装するためには経験の再現性を評価するために，抽象的な経験の表象が必要となる．これに関しては音声認識研究において提案された変換不変性を用いた抽象化を，ロボットのセンサ情報の抽象化の問題に一般化して利用する．

以下，第 2 章では変換不変性を利用したセンサ情報の抽象化について論ずる．第 3 章では抽象的知識を用いたトップダウン処理と内発的動機づけに基づく強化学習手法を提案し，第 4 章においてその有効性をシミュレーション実験によって示す．最後に第 5 章で本論文を総括する．なお，本論文は[12]の成果を詳述し，さらに発展させた内容となっている．

2. 変換不変性に基づく内発的動機づけ

ロボットの行動に伴ってセンサ-モータ系に観測されるパターンの同一性の認識は，自律ロボットに求められる重要な要件の一つである．本論文における問題設定に対しては，環境に関する明確な事前知識を用いることなく，一連の行動の結果得られるデータ系列の類似性を評価する尺度が必要となる．そこで，環境認識に関する抽象的な経験の表象としてセンサ空間（あるいは特徴空間）における変換不変量を導入する．ここでは，特に音声認識研究において Qiao らによって提案されたアファイン変換不変性[13]を用いる．変換不変性をロボットの環境認識に利用した研究としては群論的な視点からブートストラッピングの問題を扱った Censi らのもの[14]があり，本論文のアイディアと共通した部分をもつが，筆者らは経時的な表象の構造化に着目しているという点で異なっている．

$\boldsymbol{X}_{t-k_1:t+k_2} = [\boldsymbol{x}_{t-k_1}, ..., \boldsymbol{x}_t, ..., \boldsymbol{x}_{t+k_2}]$ を特徴ベクトル $\boldsymbol{x}_t \in \mathbb{R}^d$ の系列とする．$\boldsymbol{x}_t$ へのいかなるアファイン変換

$$\bar{\boldsymbol{x}}_t = \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{c} \tag{1}$$

に対しても，$\boldsymbol{X}_{t-k_1:t+k_2}$ に関するアファイン変換不変量 M は $M(\boldsymbol{X}_{t-k_1:t+k_2}) = M(\bar{\boldsymbol{X}}_{t-k_1:t+k_2})$ を満足する．ここで，$\bar{\boldsymbol{X}}_{t-k_1:t+k_2} = [\bar{\boldsymbol{x}}_{t-k_1}, ..., \bar{\boldsymbol{x}}_t, ..., \bar{\boldsymbol{x}}_{t+k_2}]$ である．本論文では，以下の形式のアファイン変換不変量を適用する．

$$M(\boldsymbol{X}_{t-k_1:t+k_2}) = \sqrt{(\boldsymbol{\mu}_a - \boldsymbol{\mu}_b)^T (\boldsymbol{\Sigma}_a + \boldsymbol{\Sigma}_b)^{-1} (\boldsymbol{\mu}_a - \boldsymbol{\mu}_b)} \tag{2}$$

$\boldsymbol{\mu}_a$，$\boldsymbol{\Sigma}_a$ はそれぞれ $\boldsymbol{X}_{t-k_1:t+k_2}$ の任意の部分列 $\boldsymbol{X}_a := \boldsymbol{X}_{t-k_1:t-1}$ の平均と共分散行列である．

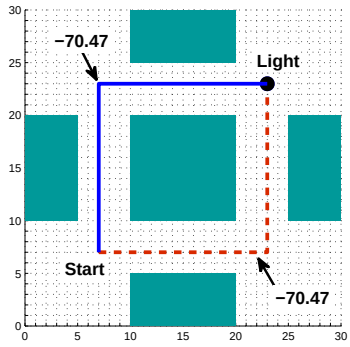
$$\boldsymbol{\mu}_a = \frac{1}{k_1} \sum_{\tau=t-k_1}^{t-1} \boldsymbol{x}_\tau \tag{3}$$

$$\boldsymbol{\Sigma}_a = \frac{1}{k_1} \sum_{\tau=t-k_1}^{t-1} (\boldsymbol{x}_\tau - \boldsymbol{\mu}_a)(\boldsymbol{x}_\tau - \boldsymbol{\mu}_a)^T \tag{4}$$

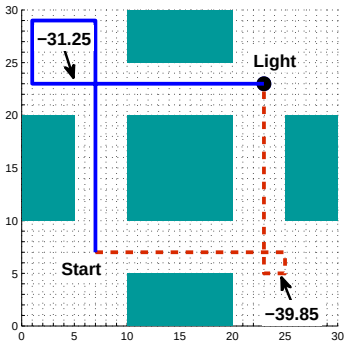
$\boldsymbol{\mu}_b$，$\boldsymbol{\Sigma}_b$ も部分列 $\boldsymbol{X}_b := \boldsymbol{X}_{t:t+k_2}$ に対して同様に定義する．音声認識研究においては，声道長や収録機器の違いはケプストラムベクトルに対するアファイン変換によって近似的にモデル化できることが知られている．つまり，アファイン変換不変量は大規模データからの話者正規化を行うことなく，ある音声言語特有の構造に対する尺度を提供する．

ロボットのセンサ-モータ系の任意の分節に対しても同様の考え方が成り立ち，ロボットと環境の相互作用の帰結として，選択された行動系列と環境の特性に依存してセンサ空間に現れるデータ系列の幾何的構造を観測することが可能である．例として，測距センサを搭載した移動ロボットが壁沿い走行を行っている状況を想定してみよう．このとき，壁沿い走行をするロボットにとって進行方向のどちら側に壁が存在するのかという情報，つまり並進運動に対して対称な情報を区別することは重要ではないはずである．ここで唯一重要なのは，壁沿い走行動作に付随してセンサ空間に観測される，壁と一定の距離を保てているか，あるいは離れてしまった場合に速やかに軌道を修正できているかといった，センサ空間においてデータ系列のなす幾何的な構造である．交差点での右折や歩行者とのすれ違いといった，より一般的な場面においても同様の議論は成り立つ．つまり，変換不変量は経時的なセンサ情報の幾何的構造から対称性等の不要な特徴を捨象した抽象的表象となり，観測された情報の同一性を計る環境認識の尺度として機能する．

実際のロボットの運用においては，しばしば高次元のセンサ空間が想定される．したがって，センサ情報から直接に変換不変量を計算するのは計算コストの面から望ましくない．加えて，アファイン変換不変量は入力ベクトルの系列がなす幾何形状に関する変換不変量であるため，センサ情報の誤差は不変量に対し大きなノイズを混入させる可能性がある．そこで前処理として，センサ情報を特徴空間に写像し低次元化を施すこととする．センサ入力 $\boldsymbol{s} \in \mathbb{R}^d$ の特徴空間への写像を $\boldsymbol{\xi} : \mathbb{R}^d \to \mathbb{R}^n$ とする．ここで $d \geq n$ である．ただし，特徴抽出によるセンサ



(a) Symmetric trajectories



(b) Asymmetric trajectories

Fig. 1　Transoformation invariance as a measure of environment recognition

情報の ξ へのコーディングはロボットの主観的な環境認識に直接的に関わるものであり，その設計の方法論に関しては画像及び音声認識等の研究における知見を参考に，今後十分な議論を行う必要がある．

図 1 では，一連の行動系列の実行の結果観測されるセンサ情報の同一性を計る尺度としてのアファイン変換不変量の特性を，測距センサと光センサを持ったエージェントが，障害物の存在する 2 次元グリッドワールド内を移動するシミュレーション実験で示している．エージェントの行動は，1 ステップに上下左右いずれかの方向に 1 セル移動する 4 種類とする．また，測距センサは上下左右の 4 方向について障害物からの距離を計測し，光センサは任意の位置に設置された一つの光源からの距離情報をエージェントに与えるものとする．図 1(a) の 2 つの経路は対称であり，エージェントは光センサから同一のセンサ入力の系列を得る．一方で，測距センサからはそれぞれの経路で異なるデータ系列を観測することになる．しかしながら，アファイン変換不変量はどちらの経路に関しても-70.47 となり，観測されたセンサ情報の系列は同一と判別される．このことは，一方のセンサ情報のなす幾何的構造が，あるアファイン変換によって他方のなす幾何的構造と一致することを示している．図 1(b) では，図 1(a) での経路にオープンスペースを見回る経路が加えられている．2 つの経路は非対称であり，見回り動作の長さによって全体の経路長も異なるものとなっている．そのため，変換不変量は長い経路で-31.25，短い経路で-39.85 と異なる値を示し，同一とは判別されない．ここで，図 1(b) における 2 経路の間の変換不変量の差は 8.60 であり，それぞれの図 1(a) における不変量との差 39.22 及び 30.62 より小さな値をとっている．したがって，見回り動作の有無は，見回り動作の短長よりアファイン変換不変量の意味では大きな差異と認められる．また，図 1(b) の短い経路は，同図の長い経路より図 1(a) の経路との類似性が高い．以上のように，センサ情報から得られる特徴ベクトルの系列間の幾何的変換に関する類似性を測る尺度として変換不変量を導入することで，時間，空間的なスケールは異なるが同一の特性をもつ情報の類似性を評価することが可能となる．

3.　経験の再現性から内発的に動機づけられた強化学習

アファイン変換不変量をスキル実行に関する環境認識の尺度とする，内発的に動機づけられた強化学習手法を提案する．ここで S をセンサ空間，A を行動空間とする．提案手法は方策オフ型 TD 学習（Temporal Difference Learning）の一つである Q-learning[(15)]の枠組みと並行に，Semi-Markov Decision Process（SMDP）における学習則が適用される．図 2 に提案手法の概略図を示す．本論文ではスキル Λ は過去に経験したタスクにおいて獲得された最適方策により得られた行動系列と，その実行に伴って観測されたアファイン変換不変量の組として定義される．以後，いくつかのスキルがそれぞれ異なる環境において事前に獲得されていることとして議論を進める．

3·1　スキルの再現性と内発的動機づけ

実行されたスキルを評価する一つの方法は，通常の強化学習と同様に外発的報酬及び価値関数である．本論文ではそれらに加えて，新たな環境においてスキルを実行した結果得られる，過去の成功経験の再現性に基づく内発的報酬によって評価を与える．環境との経時的な相互作用により得られるセンサ情報の履歴は，スキル実行の

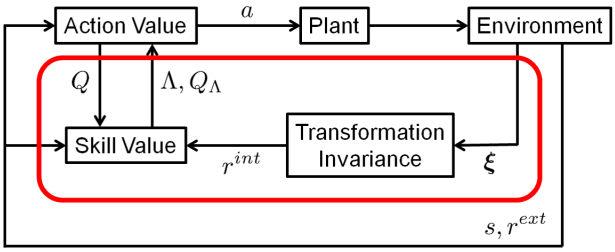


Fig. 2　Framework of proposed method

帰結であると解釈できる．スキルは過去の成功経験に基づいて構成されるため，現在のスキル実行及び過去の経験において得られたそれぞれのアファイン変換不変量で評価される状況の再現性によってスキルに対する内発的報酬を決定する．

ここで，内発的報酬 $r^{int}:\mathbb{R}\times\mathbb{R}\to\mathbb{R}$ を以下の通りに定義する．現在の環境においてスキル Λ を実行した結果得られる変換不変量を M，Λ のもつ変換不変量を M_Λ とすると，

$$r^{int}(M,M_\Lambda)=r_p^{int}\exp\left(-\frac{M-M_\Lambda}{\rho}\right) \tag{5}$$

$r_p^{int}\in\mathbb{R}_+$，$\rho\in\mathbb{R}_+$ は正のパラメータである．この定式化によりアファイン変換不変量の意味で高い再現性を示したスキル程，高い報酬が与えられることになる．その結果，過去の成功経験に対する類似性の高いスキルは学習の過程で正の強化を受けるため，現在の環境においても一貫性のある行動系列が選択されやすくなる．以下では強化学習の枠組みにおいて，上述の内発的報酬をスキル選択過程に組み込んでいる．ここで，内発的報酬それ自体はタスクとは独立であることに注意されたい．つまり，内発的報酬に基づくスキルの強化はタスクに対する効率性の向上に寄与する保証を何ら与えるものではない．

3·2　行動価値及びスキル価値の更新則

提案手法では行動価値関数 $Q(s,a)$ とスキル価値関数 $Q_\Lambda(s,\Lambda)$ が並列に学習される．ここで $s\in S$ は状態，$a\in A$ は行動である．行動価値関数は Q-learning の学習則により更新される．

$$Q(s,a)\leftarrow(1-\alpha)Q(s,a)+\alpha\left(r_{t+1}^{ext}+\gamma\max_{a'}Q(s',a')\right) \tag{6}$$

$\alpha\in[0,1]$ は学習率，$\gamma\in[0,1]$ は割引率である．$r_{t+1}^{ext}\in\mathbb{R}$ は外発的報酬，$s'\in S$ は a の実行により遷移する次ステップにおける状態である．行動価値の更新則は Q-learning と同一であり，並列に学習されるスキル価値から明示的な作用を受けない．そのため，行動選択には ε-greedy やソフトマックス行動選択[(1)]といった，Q-learning に適用される任意の方策が利用可能となっている．ただし，後述する通りスキル価値は行動選択過程における行動価値への一時的なバイアスを通して行動価値と関係づけられる．

スキル価値関数 $Q_\Lambda(s,\Lambda)$ には SMDP に拡張された Q-learning [(16)] と類似した更新則を適用する．$Q_\Lambda(s,\Lambda)$ は各スキルの行動系列の終端，またはエピソード的タスクを想定する場合はエピソードの終了条件を満足した場合にのみ更新される．

$$Q_\Lambda(s_\Lambda,\Lambda)\leftarrow(1-\alpha)Q_\Lambda(s_\Lambda,\Lambda)+\alpha\left(r^{int}+R_\Lambda+\gamma^{T_\Lambda}\max_{a'}Q(s',a')\right) \tag{7}$$

ここで，$s_\Lambda\in S$ はスキル Λ の実行が開始された状態，$T_\Lambda\in\mathbb{R}_+$ はスキルの実行開始からの経過時間を表す．$R_\Lambda\in\mathbb{R}$ はスキル実行過程において累積される，割り引かれた外発的報酬の総和であり，以下で求められる．

$$R_\Lambda=\sum_{t_\Lambda=1}^{T_\Lambda}\gamma^{t_\Lambda-1}r_{t+t_\Lambda}^{ext} \tag{8}$$

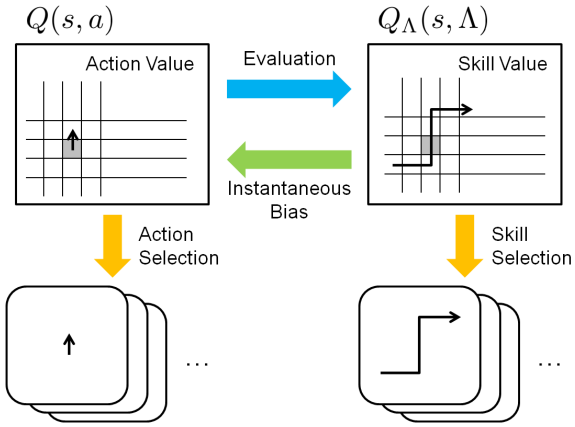


Fig. 3　Relationship between action value $Q(s,a)$ and skill value $Q_\Lambda(s,\Lambda)$

行動価値の時間的拡張に対する自然な形式としてスキル実行中の即時報酬は γ によって割り引かれた上で累積され，スキル実行が完了した時点で更新に利用される．スキル実行の結果，行動価値及び外発的報酬の意味でよりよい状態へ遷移するスキルが強化されるが，それに加えてここではスキル価値の更新において内発的報酬 r^{int} が用いられる．そのため，スキル実行の間に観測されるセンサ情報の系列が，そのスキルに紐付いた過去の成功経験に対し，アファイン変換不変量を尺度として高い再現性を示す場合，そのスキルはタスクとは独立に報酬が割り当てられ，再選択が動機づけられる．

3·3　スキルによる行動へのバイアス

option の枠組み等と異なり，提案手法では時間的に拡張された行動であるスキルと 1 ステップの行動は区別して扱われる．行動価値はスキル価値によるバイアスを加えられ，スキルは間接的に行動価値に埋め込まれる．スキル価値の更新においては，スキル実行結果の評価に T_Λ ステップに渡る行動価値の TD 誤差[(1)(16)]が用いられる一方で，行動価値の更新則にスキル価値は直接的には利用されない．つまり，選択されたスキルの実行に沿って，各時刻でスキルが指定する行動 a の価値 $Q(s,a)$ はスキル価値 $Q_\Lambda(s_\Lambda,\Lambda)$ に基づくバイアスを受ける．ある時刻において実行中のスキルの指定する行動の価値は以下の通り変更される．

$$Q(s,a_\Lambda) \xleftarrow{bias} Q(s,a_\Lambda) + \beta\gamma^{-(t_\Lambda-1)}Q_\Lambda(s_\Lambda,\Lambda) \tag{9}$$

t_Λ はスキル Λ が選択されてからの経過時間，$a_\Lambda \in A$ は t_Λ においてスキルの指定する行動，$s_\Lambda \in S$ はスキルの実行が開始されたときの状態である．$\beta \in \mathbb{R}$ は a_Λ に対するバイアスの強度を決定するパラメータである．a_Λ に対するバイアスは各時刻において，割増されたスキル価値によって決定される．式 (9) は更新則ではなく，行動価値に対する一時的な加算であり，バイアスが加えられる行動価値 $Q(s,a_\Lambda)$ は，その時刻における行動が選択された時点で元の値に戻される．

$Q(s,a)$ と $Q_\Lambda(s,\Lambda)$ の関係性は図 3 の通りである．スキル価値は式 (7) にしたがって，過去の経験の再現性に基づく内発的報酬，外発的報酬及び行動価値によって更新される．他方，行動価値の更新にスキル価値は用いられない．しかし，スキルは式 (9) によって行動選択過程におけるバイアスとして作用する．選択されたスキルの価値が高いものであった場合，各時刻においてスキルの指定する行動は選択される確率は高くなる．行動価値は式 (7) によって更新されるため，その行動が TD 誤差の意味でよりよい状態へエージェントを遷移させたならば，スキルは行動価値上に埋め込まれることになる．逆にスキルの指定する行動がよくないものであれば，その状態においては別の行動が選択されやすくなり，スキルは修正された形で行動価値上に表現される．以上のように，行動価値の学習による修正された形でのスキルの埋め込みとスキル選択の学習が並列に行われる．

3·4 適格度トレースとスキルの効用問題

スキルはある特定の環境における最適な行動系列から定義される．未知環境での学習において有効にスキルを利用するためには，様々な状況を記述可能な多様性を担保する，十分な数のスキルを予め獲得しておくことが望ましい．しかしながら，このときスキル選択過程の肥大化による学習の減速が起こり得る．また，スキルは行動を時間的に拡張するため，誤った選択によるパフォーマンスの低下は 1 ステップの行動と比較して大きくなってしまう．この問題に対するアプローチの一つは，環境に対応する有用なスキルの選択方策というメタ知識の探索であるが，これはマクロオペレータによる効率化学習における効用問題[17]と等価の困難さをもつ．

そこで，環境情報から適切なスキルを選択するための知識の蓄積ではなく，行動価値へのスキルの埋め込み及びスキル価値の学習過程，言い換えればトップダウン処理による概念化の加速を行うために適格度トレースを導入する．行動価値 $Q(s,a)$ に関しては以下の形式の入替え更新トレースを用いた naive Q(λ) を適用する．

$$e(s,a) \leftarrow \begin{cases} 1 & (s = s_t, a = a_t) \\ 0 & (s = s_t, a \neq a_t) \\ \lambda \gamma e(s,a) & (s \neq s_t) \end{cases} \tag{10}$$

ここで $s_t \in S$ 及び $a_t \in A$ は現在の状態及び行動である．スキル価値 $Q_\Lambda(s,\Lambda)$ に関しては以下のトレースを用い，スキル価値同様にトレースの更新はスキルの終端において行う．

$$e_\Lambda(s,\Lambda) \leftarrow \begin{cases} 1 & (s = s_{t_\Lambda}, \Lambda = \Lambda_{t_\Lambda}) \\ 0 & (s = s_t, \Lambda \neq \Lambda_{t_\Lambda}) \\ \lambda \gamma^{T_\Lambda - 1} e_\Lambda(s,\Lambda) & (s \neq s_{t_\Lambda}) \end{cases} \tag{11}$$

$s_{t_\Lambda} \in S$ はスキルの開始された状態，Λ_{t_Λ} は実行されたスキルである．適格度トレースを導入することで，現在の行動価値及びスキル価値に関する TD 誤差が過去の経験にバックトラックされるため，スキル価値に基づくバイアスによるスキルの行動価値への埋め込み及びスキル選択の評価が加速する．これはトップダウン処理による環境の概念化を推し進めるものであり，因果関係を記述する範囲の拡大と解釈できる．したがって，適格度トレースによって適切なスキルとそうでないスキルの区別が高速化し，効用問題による学習性能の低下を抑えることが可能となる．

提案手法の擬似コードを図 4 に示す．基本的なアルゴリズムの構造は[12]と同一である．ただし，ここでは行動価値及びスキル価値それぞれに対する適格度トレースをそれぞれ式 (10) 及び式 (11) の形式で導入している．また，2 次元グリッドワールドにおけるナビゲーション問題に提案手法を適用した結果得られた経路と選択されたスキルの例を図 5 に示す．初期状態を $(1,1)$ とし，目標状態 $(15,15)$ において正の外発的報酬を得ることをエピソードの終了条件とした．環境内には障害物が設置してあり，これらとの衝突を回避する最短経路を学習した．実線は実際に学習された経路であり，星はスキルが開始された状態，破線は選択されたスキルの指定する行動に従った場合の経路である．ただし，破線については障害物との衝突を無視して表示している．選択された経路の多くが，スキルの指定する行動によって得られる経路と一致していることが見て取れる．また，一部実際の経路とスキルの指定する経路が対称に近い形状をとっている部分がある．これらは行動価値の更新によるスキルの未知環境への適応及び図 1 に示したアファイン変換不変量の特性に基づく環境認識の結果である．

4. シミュレーション実験

これまで内発的に動機づけられた強化学習手法は，主にボトムアップ処理によって有用なスキルを階層的に獲得するという観点から提案されてきた．しかし，第 1 章でも述べた通り，経時的な環境との相互作用を続けスキルを蓄積し，階層構造が拡張されるにしたがって，見かけ上学習が遅くなるという問題が想定される．したがって，何らかの手段で探索空間の積極的な縮減を行う必要がある．本章ではトップダウンな知識に基づく探索戦略を Q-learning の枠組みに組み込むことで得られる効果をシミュレーション実験によって検証する．提案する手法

initialize $Q(s,a)$, $Q_\Lambda(s,\Lambda)$, $e(s,a)$, $e_\Lambda(s,\Lambda)$
repeat
  get sensor data and calculate $\boldsymbol{\xi}$
  bias $Q(s,a_\Lambda)$ instantaneously by (9)
  select a using an action selection policy w.r.t. biased $Q(s,a)$
  observe r^{ext}_{t+1} and next state s'
  update $e(s,a)$ by (10)
  update $Q(s,a)$ and R by (6), (8)
  if $t_\Lambda = T_\Lambda$ **then**
    calculate M and r^{int} by (2) and (5)
    update $e_\Lambda(s,\Lambda)$ by (11)
    update $Q_\Lambda(s_\Lambda,\Lambda)$ by (7)
    select next skill Λ' using a skill selection policy w.r.t $Q_\Lambda(s',\Lambda)$
  end if
  $t \leftarrow t+1$, $t_\Lambda \leftarrow t_\Lambda+1$, $s \leftarrow s'$
until reach destination state or termination time
update $e_\Lambda(s,\Lambda)$ by (11)
update $Q_\Lambda(s_\Lambda,\Lambda)$ by (7)

Fig. 4　Summary of algorithm

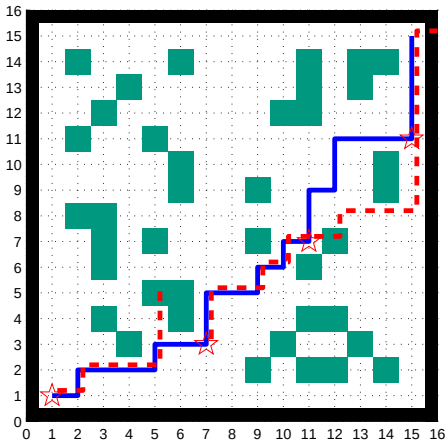


Fig. 5　Resulted path and selected skills by proposed method

は SARSA などの TD 学習手法や，価値関数転移手法との併用も可能である．そのため見通しのよい実験結果を示すため，タスクはシンプルな 2 次元グリッドワールド環境におけるナビゲーション問題を扱うこととする．また，同様の理由から提案手法の基盤となっている Q-learning または naive Q(λ) をベースラインとして実験を行う．まず，適格度トレースを用いない提案手法によって，適切なスキルが獲得された場合には変換不変性を利用した内発的動機づけによって学習が加速可能であり，さらに状態数の増大に対して安定した学習を行うことができるという提案手法の特性を示す．また，適格度トレースの導入と多様なスキルに対する提案手法の応答について検証する実験を行い，提案手法の一般的な問題への適用について議論する．

実験設定は以下の通りである．エージェントの行動は上下左右方向のいずれかに 1 セル移動する 4 種とした．環境は壁に囲まれた正方形領域とし，1 セルを占有する 1 辺が 1 の長さの正方形障害物を複数設置した．ただし，初期

状態と目標状態を結ぶ線分を対角線とする正方形領域を囲む壁面を設置し，壁と隣接するセルには障害物は設置しない．エージェントの初期状態を $(1,1)$，目標状態を s_d とし，s_d においてエージェントは 5 の外発的報酬を受け取り，障害物に衝突した場合は-1 の報酬を受ける．また，各行動には-0.1 の移動コストがかかることとした．$Q(s,a)$ 及び $Q_\Lambda(s,\Lambda)$ は初期状態で $[0,1]$ の間の一様分布から各状態についてランダムな値を設定した．$\rho=10$，$r_p^{int}=1$，$r_n^{int}=0.5$，$\alpha=0.2$，$\gamma=0.95$，$\beta=0.5$，エピソードの終了条件は s_d への到達，もしくは一定の時間ステップ数の経過とした．行動選択には以下のソフトマックス行動選択を適用した．スキル選択に関しても同様である．

$$\pi(s,a)=\frac{\exp(Q(s,a)/\tau)}{\sum_{b\in A}\exp(Q(s,b)/\tau)} \tag{12}$$

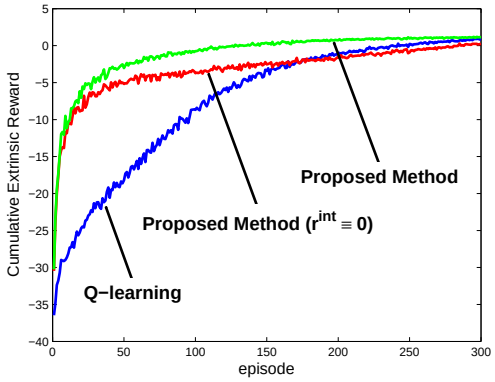
温度定数 τ は行動及びスキル，双方について現在のエピソード数を E として，$0.2\exp(-E/150)$ とした．エージェントは上下左右 4 方向それぞれについて 2 セル先までの障害物との距離を計測できるものとし，それらの計測値の和及び s_d と現在の状態の差のユークリッドノルムを要素にもつ 2 次元ベクトルの時間微分を特徴ベクトル $\boldsymbol{\xi}$ とした．アファイン変換不変量の計算において，$\boldsymbol{\xi}$ の系列は中間地点で分割した．スキルは各々の新規の環境について構成し直した．スキルを獲得する環境における s_d は $(5,5)$ で，5 個の障害物を設置し，Q-learning によって最適方策を求めた．

4·1 内発的動機づけとスキルによる探索空間の縮減

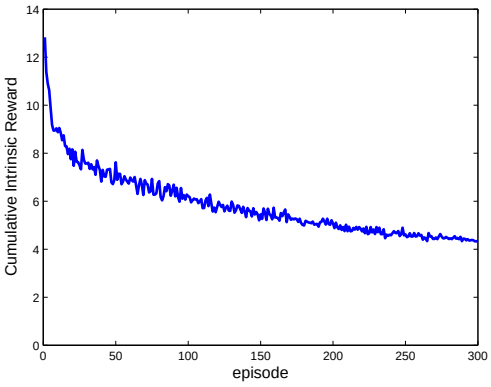
アファイン変換不変量に基づいて与えられる内発的報酬の学習過程への効果を調べるために，適格度トレースを用いない提案手法，内発的報酬を $r^{int}\equiv 0$ とした提案手法，Q-learning のそれぞれについて実験を行った．45 点の障害物を設置し，スキル数は 5 とした．また，$s_d=(15,15)$，エピソードの終了時間は 300 ステップとした．障害物の配置，スキルのそれぞれ異なる 15 試行の実験結果の平均を図 6 に示す．それぞれの試行について，300 エピソードの実験を 10 回行った結果の平均を用いている．図 6(a) はそれぞれの手法の外発的報酬に関する学習曲線を示している．縦軸はエピソード毎の獲得報酬の総和，横軸はエピソードである．行動価値の更新則自体は 3 つの手法で同一であるが，学習曲線には顕著な違いがみられる．この結果は学習初期における立ち上がりと，その後の収束の二点にわけて考えることができる．

まず，開始から 20 エピソード程度の学習の初期段階において，提案手法と内発的報酬を与えない提案手法は Q-learning より速い立ち上がりを示した．スキルがバイアスを与える行動系列 A_Λ は少なくとも一つの環境における最適方策から抽出されている．そのため成功経験として定義されたスキルを用いることで，実際に選択される一連の行動に一貫した動作戦略が反映される．その結果狭められた探索空間において学習が行われ，行動価値上にスキルが埋め込まれることで目標状態へ至る大雑把な経路が構成されることになる．

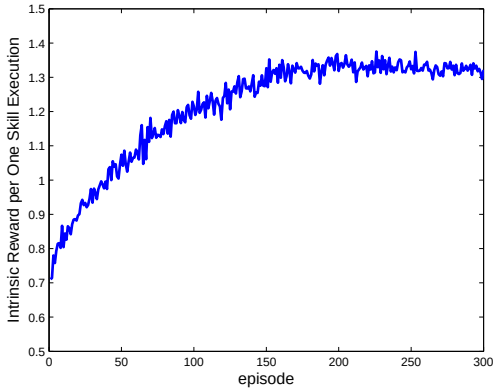
しかし，30 エピソード程で内発的報酬を与えない提案手法は学習が停滞し始め，120 エピソード程で Q-learning の方が高い累積報酬を獲得するようになってしまっている．一方で内発的報酬を導入した場合は学習が進行し続け，最適方策に収束している．二手法の差は内発的報酬の有無のみであり，この結果はそれを反映している．上述の通り，スキルは少なくとも一つの特定の環境における最適方策としての整合性をもつため，スキルの利用により大雑把な目標状態への経路を，探索の初期段階において獲得することはある程度可能である．しかしながら，スキルによるバイアスは学習を阻害することもある．行動価値は MDP において学習される一方，スキル価値は時間的に拡張された SMDP において学習されるため，行動価値と比較して更新される頻度が低い．したがって，行動価値とスキル価値を並列に学習するにあたって，それらの間に（単純に考えれば 1 回のスキル実行にかかる時間分の）学習速度の差異が生じる．そのため，行動価値が正しい行動を選択するよう学習されていたとしても，そのような行動選択を阻害するようなバイアスが加えられるという問題が起こり得る．そこで，内発的報酬，すなわち変換不変性に基づく環境認識によるスキル選択の評価と，それによる行動選択過程に対するバイアスの調整が重要となる．選択されたスキルに対して，それがどの程度整合性のある経験を生んだのかを内発的報酬によって評価することで，行動価値へのバイアスが制御される．その結果，行動価値及び報酬に関してよりよい状態へ遷移し，かつ一貫性のある一連の状態行動対がより強く強化されることで探索空間の縮減が行われ収束性能が向上する．さらに，利用できるスキルが環境内のある状態に対して適合するものでなかったとしても，内発的報酬



(a) Learning curve of extrinsic reward



(b) Learning curve of intrinsic reward



(c) Averaged intrinsic reward per one skill execution

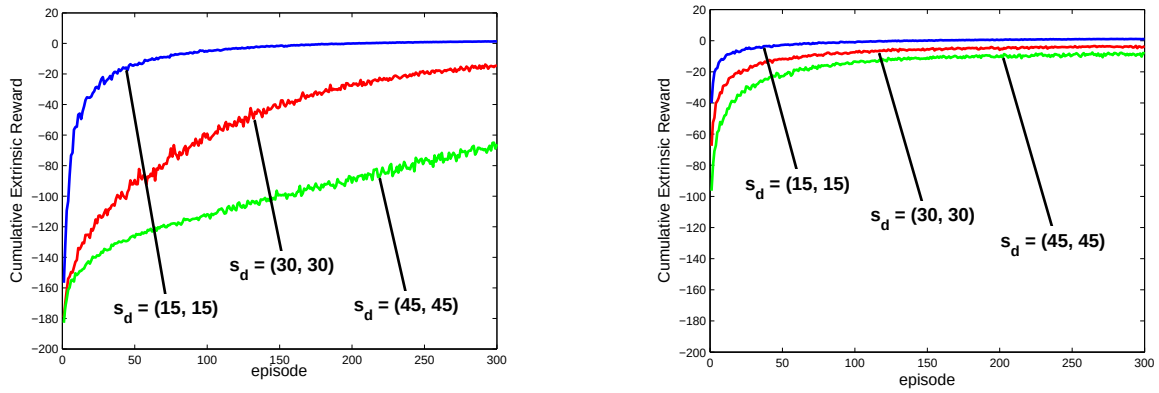
Fig. 6 Effect of intrinsic reward via transformation invariance

により行動価値の更新に基づくスキルの環境に対する選択的な適応がなされ，修正されたスキルの評価によってトップダウンなみなしは期待される形に環境の概念化を積極的に推し進めることが可能となる．

図 6(b) は内発的報酬の遷移を示しおり，縦軸はエピソード毎の獲得報酬の総和，横軸はエピソードである．学習初期においては高い内発的報酬を獲得しているが，これは探索のために多くのスキルを利用した結果である．20 エピソード程度までに学習が急速に進み，以降も図 6(a) で示される学習の進行と同期して，獲得される内発的報酬は漸減した．図 6(c) は，1 つのスキル実行あたりの内発的報酬のエピソード毎の平均である．図 6(b) における累積報酬と対照的に，スキルあたりの内発的報酬は学習の進行に対して増加傾向をもつ．スキルの実行は必ずしも学習のパフォーマンスを向上するものではない．しかし，学習の初期段階においては不適切なスキルの価値に基づいて行動に負のバイアスが加えられ，スキルが否定的な概念として働くことでも学習が効率化される．その結果，学習が進むにつれてアファイン変換不変量の意味で過去の成功経験をよく再現するスキルが選択的に実行されるようになっている．また，図 6(c) において，約 250 エピソード目から内発的報酬がわずかに減少傾向を示しているのは，スキルによって大まかに形成された目標状態への経路が，行動価値の更新によるボトムアップ処理で修正された結果であると考えられる．

次に，目標状態 s_d を $(15,15)$，$(30,30)$，$(45,45)$ に設定した 3 つの環境について実験を行い，状態数を上述の実験からそれぞれ 4 倍，16 倍に増やした環境について検証した．障害物数はそれぞれの環境について 45，180，405 とし，エピソードの終了時間は 1500 ステップ，スキル数は 5 とした．障害物の配置，スキルのそれぞれ異なる 15 試行の実験結果の平均を図 7 に示す．それぞれの試行について，300 エピソードの実験を 10 回行った結果の平均を用いている．図 7(a) 及び 7(b) はそれぞれ Q-learning 及び提案手法の学習曲線である．縦軸はエピソード毎の獲得報酬の総和，横軸はエピソードである．

状態数の増大に対して，Q-learning では学習曲線の形状に違いが表れている．初期条件やパラメータ設定にも依



(a) Response of Q-lerning

(b) Response of proposed method

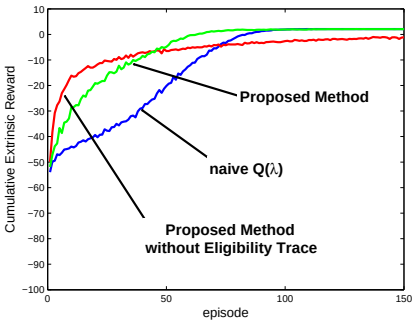
Fig. 7 Response to scaled environment

存するが，一定の条件の基で大域的最適性を保証する Q-learning は探索空間全体を偏りなく探索する傾向があるため，図 7(a) のように探索空間の増大に伴う性能の低下が起こる．それに対して，提案手法では探索空間の増大に対する学習性能の低下がほとんど起こらず，各々の環境に対して学習曲線の形状に大きな違いはみられない．提案手法においても，行動価値は Q-learning と同一の更新則にしたがっている．したがって，この結果は提案手法におけるトップダウンなスキルに基づく選択的な探索がなされたことを意味する．探索戦略が一貫しており，かつ目標状態へ至る経路の固定化を，スキル価値を用いて行うため，状態数に依存せず安定した学習を行うことが可能となる．指向性のある探索と，スキル価値に対する内発的報酬による探索空間の絞り込みは学習の初期段階で特に大きな効果を発揮する．未学習の段階では価値関数に大きな勾配はないため，Q-learning では初期状態を中心に等方的に探索範囲が広がっていくが，提案手法は高い再現性を示したスキルによって形成される経路を中心とした探索が行われる．そして，そのような探索中心の候補は外発的報酬系によってさらに選別される．そのため，提案手法では状態数が増大しても，それに伴って増加する学習コストはほとんど探索中心の構成及び経路の調整程度であるため，安定した学習結果を得ることができる．

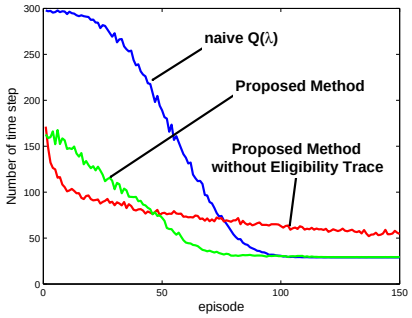
4·2 スキルの効用問題

これまでの実験において利用されたスキル（以後，適切なスキルと呼ぶ）は，テストされる実験環境をスケールダウンした環境において学習された最適方策から抽出されており，外壁を含む障害物の配置は異なるものの，実験環境においても有効に作用する可能性をもったものが用いられていた．しかし，本論文では事前条件を用いない未知環境における学習過程を想定しているため，多様な環境に対応する十分な種類のスキルを事前に獲得し，さらにそれらを有効に利用できることが望ましい．そこで本小節では，これまでの実験よりも多様な環境において獲得されたスキル群を用いて，適格度トレースを導入した提案手法について実験を行った．したがって，選択可能なスキルの中にはそれに従った場合，明らかに現在の実験環境におけるパフォーマンスを低下させるもの（以後，不適切なスキルと呼ぶ）が混入することになる．実験では，4·1 節同様のスキル設定に加えて，それぞれの初期状態 $s_{o_1}=(1,1)$，$s_{o_2}=(1,5)$，$s_{o_3}=(5,1)$，$s_{o_4}=(5,5)$ に対し，目標状態 $s_{d_1}=(5,5)$，$s_{d_2}=(5,1)$，$s_{d_3}=(5,1)$，$s_{d_4}=(1,1)$，$s_d=(15,15)$ と設定した環境において 5 ずつ，計 20 のスキルを獲得した場合及び 20 ずつ，計 80 のスキルを獲得した場合の 2 つの条件での実験を行った．スキル獲得環境に関する他の設定はこれまでと同様である．実験環境の障害物数は 45，エピソードの終了時間は 300 ステップ，$r_p^{int}=1$，$\lambda=0.8$ とした．提案手法，適格度トレースを用いない提案手法及び naive Q(λ) の各手法について，障害物の配置，スキルのそれぞれ異なる 30 試行の実験を行った．それぞれの試行について，150 エピソードの実験を 10 回行った結果の平均を用いている．各々のスキル獲得環境の設定に関する実験結果を図 4·2 及び図 9 に示す．

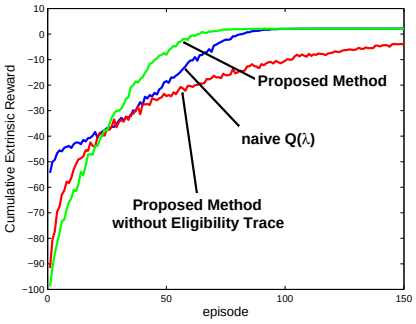
図 8(a) 及び図 8(b) はこれまでの実験と同様の，s_{o_1} と s_{d_1} の組について獲得した 5 つの適切なスキルのみを用いた場合の結果で，それぞれエピソード毎の外発的報酬の総和及びエピソード終了までの経過時間ステップある．トレースを用いることで学習初期の性能は低下するものの，最適方策への収束は早くなることが確認された．これ



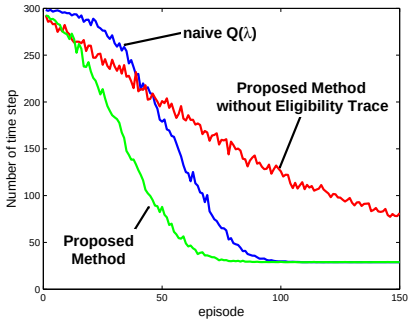
(a) Learning curve of extrinsic reward by appropriate 5 skills



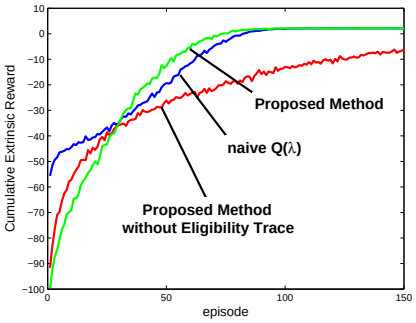
(b) Spent time step by appropriate 5 skills



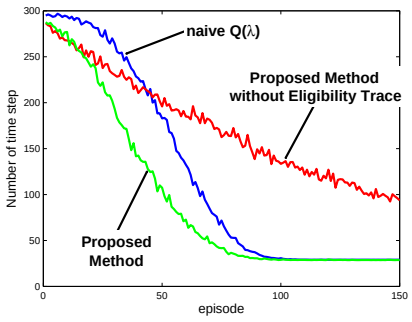
(c) Learning curve of extrinsic reward by inappropriate 20 skills



(d) Spent time step by inappropriate 20 skills



(e) Learning curve of extrinsic reward by inappropriate 80 skills



(f) Spent time step by inappropriate 80 skills

Fig. 8 Effect of eligibility trace for utility problem in skill selection process

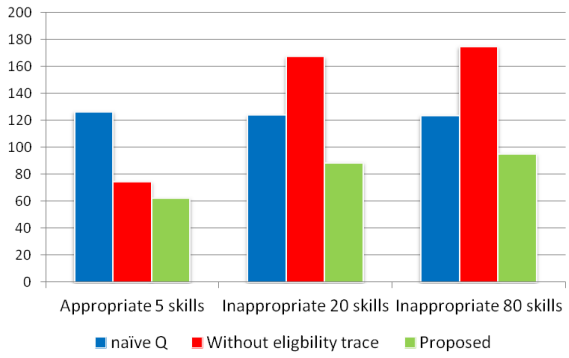


Fig. 9 Mean of cumsumed time step

はトレースによる行動，スキル双方に関する経験のバックアップを利用することで，スキルの環境への適応がより積極的に進められたためである．学習初期においてトレースを用いない方が学習が早いのは，探査的な行動が多く選択され行動価値間の相対的な差が頻繁に入れ替わるため，トレースによって誤った事象間の関連付けが行われ易いことが理由である．図 9 はエピソード当たりの終了条件を満足するまでに経過した時間ステップの平均である．全体としてはトレースの導入によって学習性能が向上することが確認できる．

次に，実験環境に対して不適切なスキルを含む 20 のスキルを獲得させた場合の結果を図 8(c) 及び図 8(d) に示す．それぞれ，エピソード毎の外発的報酬の総和及びエピソード終了までの経過時間ステップに関する結果である．この実験ではこれまで同様の 5 つのスキルに加えて，最適方策では選択されない行動を含む 15 の不適切なスキルが利用されている．したがってスキル数及びスキルの多様性の増大に対して，図 8(a) 及び図 8(b) の結果からの性能低下をどの程度抑えられるかがここでの問題である．まず，提案手法及びトレースなしの提案手法の両方で学習初期において獲得報酬の低下，必要時間ステップの増加が確認された．目標状態から必ず遠ざかるような行動をもったスキルによって探索が実行されるため，適切なスキルのみを用いる場合より探索の指向性は低下し，初期においてはこれまでと比較して等方的な探索戦略がとられる．その結果，目標状態へ至る経路候補の固定にこれまでより時間がかかり，初期の性能低下が引き起こされた．特にスキルは行動を時間的に拡張するため，誤選択によるロスが大きい．その結果，トレースを用いない提案手法は 150 エピソード以内に収束しない程に性能が低下した．他方で，トレースを用いた場合には学習初期の性能低下は起こるものの，収束に必要なエピソード数自体はほとんど変化しなかった．図 9 をみても，トレースを用いることで学習時間の増加はかなり抑えられており，ベースラインの naive Q(λ) からの性能向上は保たれている．トレースを導入することで，スキルの誤選択によって実行された行動群の学習過程が時間的に拡張される．学習初期の比較的フラットな価値関数に関する，ある状態における一回の不適切なスキルの実行を考えてみよう．スキル実行のプロセスで選択された一連の行動の結果観測されるそれぞれの TD 誤差は様々な値となり得る．しかし，スキル全体としては不適切なスキル実行の結果，負の TD 誤差を観測することになりスキル価値は減少すると考えられる．また，トレースを導入することで現在の価値観数の更新が過去にさかのぼって行われるため，一連の行動状態対の価値も 1 ステップ更新の場合と比較して，一様に低い値をとることになる．その結果，次の状態の訪問において一度実行された不適切なスキルが選択される可能性は低下し，仮に選択されたとしても負のバイアスによる探索空間の縮減が起こり学習を進行させることが可能となる．この一連の過程により，初期の学習が遅延するものの，大きなパフォーマンスの低下を起こすことなく，内発的報酬による経路候補の絞り込みが行われ全体としての学習性能の向上効果を得ることができる．

ただしこの方法にも限界がある．スキル数を 80 に増やした場合の図 9 の結果をみると，スキル数が増えることでも性能の低下が起こることが確認できる．エピソード毎の外発的報酬の総和及びエピソード終了までの経過時間ステップに関する結果を示した図 8(e) 及び図 8(f) を図 8(c) 及び図 8(d) と比較すると，不適切なスキルの混入による初期の学習性能の低下とは異なり，収束時間の増加を含め，全体的にパフォーマンスが低下していることが確認できる．大きな要因としては行動選択に soft-max 選択を用いているため，スキル数が増える程に相対的に価値関数の差異が小さくなり，様々なスキルが実行されやすくなる．すると選択的な探索戦略という提案手法の特徴が十分に機能しなくなる．つまり，スキル価値の学習過程の非効率化が問題となる．行動選択を ε-greedy などによって行えば部分的にはこの問題は解決すると考えられる．この問題に限らず機械学習的な観点からみれば提案手法はむしろ ε-greedy と相性がよい場合もある．だが，ロボットの発達過程においてトップダウン処理と ε-greedy のような形式での集中的な行動の選択方式を用いることは，強すぎる環境の概念化をもたらすことになると考えられる．つまり，本論文の文脈で言えば経路候補が極めて少なく，何らかの環境の変化によってそれが利用できなくなった場合の復帰に困難を伴う非適応的な学習システムとなってしまう可能性がある．そのために本論文では soft-max 選択を適用しており，したがって，スキル数の増大に関しては，今後さらなる検討が必要であると考えられる．

5. 結　　言

本論文ではトップダウン処理のもつ，環境情報の知識による積極的な概念化という機能に着目し，過去の成功経験の新たな環境における再現という内発的動機づけに基づく強化学習手法を提案し，その有効性を検証した．提案

手法では過去の成功経験における行動系列と，それに伴って観測されるセンサ情報の系列を抽象化した知識の組であるスキルを用い，MDP 及び SMDP における学習が相互に作用しながら並列に実行される．選択された行動系列の帰結である，特徴空間に観測されるデータ系列に関する幾何的な変換不変量を環境認識の尺度として導入し，これにより特定の状態空間に依存しないスキルに対する評価を行う．高い再現性を示したスキルは外発的報酬と行動価値に加えて，内発的報酬によって動機づけられる．また，スキルはその価値に基づいて，内発的報酬によって制御されるバイアスを行動価値に与え，行動価値上に適応的に埋め込まれる．スキルの再現性に基づく内発的動機づけはタスクと独立しており，新たな環境において妥当な行動を与える保証を与えるものではない．しかし，成功経験から抽出されたスキルを再現することで，探索と利用の戦略に一貫性が生じ探索空間が狭められる．その結果，環境情報の知識に基づく概念化を積極的に促進することで学習が加速することをシミュレーション実験によって示した．また，状態数の変化に対しても提案手法の選択的な探索空間の構造化が有効に働き得ることを示した．スキルの効用問題に対しては，行動価値及びスキル価値それぞれに関する適格度トレースを導入することで，多様なスキルを用いることによる性能の低下を抑えることが可能であることを確認した．ただし，スキル数の増大に伴う性能低下に関しては十分に解決されておらず，今後さらなる検討が必要である．

トップダウン処理によるみなしは環境の概念化を推し進める強力なツールになり得るが，ボトムアップ処理もまた適応的かつ自律的なロボットの構成に重要である[18]．例えば，本来スキルの獲得は連続的なエージェントの学習過程において自律的に獲得されることが望ましいが，感覚運動系における時系列の分節化やモデルに基づく制御のための予測器の構築においてはボトムアップ処理が必要となる．したがって，トップダウン処理とボトムアップ処理が相補的に機能する包括的な発達的システムの構築は今後の課題である．そのためには行動空間における抽象化によって，2 つの抽象的表象間の関係性を論ずる必要が生じるが，これに関しては統計的な観点からのアプローチが必要であると考えている．また，提案したトップダウン処理に基づく探索戦略の実装は，例えば SARSA[1]などの他の TD 学習手法，Profit sharing[19]などの経験強化型の強化学習手法，あるいは転移学習手法[20]など，Q-learning 以外の学習手法との組み合わせによる性能の向上や機械学習的な観点からの議論も今後の研究課題である．

文　　献

(1) Sutton, R. S., Barto, A. G., *Reinforcement Learning:An Introduction* (1998), Cambridge, MA, MIT Press.

(2) Weng, J., McClelland, J., Pentland, A., Sporns, O., Stockman, I., Sur, M., Thelen, E., “Autonomous Mental Development by Robots and Animals”, *Science*, Vol. 291, No. 5504 (2001), pp. 599-600.

(3) Singh, S., Barto, A. G., Chentanez, N., “Intrinsically Motivated Reinforcement Learning”, *Proceedings of the Advances in Neural Information Processing Systems 17* (2005), pp. 1281-1288.

(4) Stout, A., Konidaris, G. D., Barto, A. G., “Intrinsically Motivated Reinforcement Learning: A Promising Framework for Developmental Robot Learning”, *In AAAI Spring Symposium on Developmental Robotics*, (2005), pp. 1281-1288.

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

(6) Oudeyer, P. -Y., Kaplan, F., Hafner, V. V., “Intrinsic Motivation Systems for Autonomous Mental Development”, *IEEE Transactions on Evolutionary Computation*, Vol. 11, No. 2 (2007), pp. 265-286.

(7) Vigorito, C. M., Barto, A. G., “Intrinsically Motivated Hierarchical Skill Learning in Structured Environments”, *IEEE Transactions on Autonomous Mental Development*, Vol. 2, No. 2 (2010), pp. 83-90.

(8) Konidaris, G. D., Barto, A. G., “Building Portable Options: Skill Transfer in Reinforcement Learning”, *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, Vol. 2 (2007), pp. 895-900.

(9) 甘利俊一, 中川聖一, 鹿野清宏, 東倉洋一, 音声・聴覚と神経回路網モデル (1990), オーム社.

(10) Summerfield, C., Egner, T., “Expectation (and Attention) in Visual Cognition”, *Trends in Cognitive Sciences*, Vol. 13, No. 9 (2009), pp. 403-409.

(11) Kunda, Z., “The Case for Motivated Reasoning”, *Psychological Bulletin*, Vol. 103, No. 3 (1990), pp. 480-498.

(12) 増山岳人, 山下淳, 淺間一, “変換不変性と内発的動機づけに基づく強化学習”, 第 30 回日本ロボット学会学術講演会予稿集, (2012), AC4F1-1.

(13) Qiao, Y., Suzuki, M., Minematsu, N., “Affine Invariant Features and Their Application to Speech Recognition”, *Proceedings of the 34th IEEE International Conference on Acoustics, Speech and Signal Processing*, (2009), pp.4629-4632.

(14) Censi, A., Murray, R. M., “Uncertain Semantics, Representation Nuisances, and Necessary Invariance Properties of Bootstrapping agents”, *Proceedings of the 1st Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics*, Vol. 2 (2011), pp. 1-8.

(15) Watkins, C. J. C. H., Dayan, P., “Q-learning”, *Machine Learning*, Vol. 8 (1992), pp. 279-292.

(16) Barto, A. G., Mahadevan, S., “Recent Advances in Hierarchical Reinforcement Learning”, *Discrete Event Dynamical Systems: Theory and Applications*, Vol. 13 (2003), pp. 341-379.

(17) Minton, S., “Quantitative Results Concerning the Ttility of Explanation-Based Learning”, *Artificial Intelligence*, Vol. 42, No. 2-3 (1990), pp. 363-391.

(18) Bonarini, A., Lazaric, A., Restelli, M., Vitali, P., “Self-Development Framework for Reinforcement Learning Agents”, *Proceedings of the 5th International Conference on Development and Learning*, (2006).

(19) Grefenstette, J. J., “Credit Assignment in Rule Discovery Systems Based on Genetic Algorithms”, *Machine Learning*, Vol. 3 (1998), pp. 225-245.

(20) Taylor, M. E., Stone, P., “Transfer Learning for Reinforcement Learning Domains: A Survey”, *Journal of Machine Learning Research*, Vol. 10 (2009), pp. 1633-1685.