階層型強化学習における スキル価値の汎化に基づいた探索空間の縮減

○増山 岳人 (中央大学)　山下 淳 (東京大学)　淺間 一 (東京大学)　梅田 和昇 (中央大学)

1. 緒言

人の生活空間のような，詳細な事前知識を想定することが困難な環境へとロボットの活動領域を拡大するためには，未知の状況に自律的に適応する学習機能が必要となる．そのような要求に応える枠組みの 1 つに，強化学習 [1] が挙げられる．強化学習では，試行錯誤を通じた環境との相互作用とタスクを記述する報酬信号から，価値関数と呼ばれる行動選択の優先度を決定する関数を教師なしで学習する．試行錯誤的な行動選択と状態遷移から，タスクを達成する最適な制御方策をオンラインで獲得するため，環境や制御対象に関する詳細な情報を必要としないという利点がある．しかし，経験の蓄積から行動を学習するため，特に高次元の問題空間においては最適方策を学習するために膨大な探索行動が必要になるという原理的な問題がある．

この問題に対処するため，近年，階層化された方策によって学習を効率化する階層型強化学習 [2] が注目されている．階層型強化学習では，短時間的な行動（primitive action）だけでなく，複数の行動の組み合わせによって構成されるスキル（skill，temporally extended action，option 等と呼ばれる）を用いる強化学習の枠組みである．時定数の異なるスキルを適切に利用することで，意思決定過程を短絡し，学習を効率化することが可能となる．階層型強化学習におけるスキルは多くの場合，問題となる状態空間の部分空間における最適方策から抽出される．スキル自体が満足する最適性を前提に，スキル利用について最適な方策（recursively optimal policy）を学習することが目的とされる [3]．獲得するスキルを徐々に複雑化させる累増的なスキル学習 [4] によって，様々な場面において適応的に振る舞うロボットシステムの実現が期待されている．

しかし，スキル獲得過程における最適化基準にしたがって構成される手法には，スキルの複雑化に伴う見かけ上の学習速度の低下という問題が生じると考えられる．このような問題を Der らは次元の呪いの問題の一種であると指摘しており [5]，学習を行う状態（行動）空間が複雑である程，顕著なものとなる．あらゆる状況に対処することが可能な能力の獲得は重要である一方，人の生活空間においてタスクを実行するロボットには，最適ではなくともタスクを実行可能な解を素早く学習する能力が要求される．すなわち，まずは準最適な解を素早く学習し，その後に探索範囲を広げることで最適解を学習するという戦略が望ましいと考えられる．

2. 経験の再現性に基づく階層型強化学習

筆者らは経験の再現性に対する内発的動機づけによって，能動的に探索空間を縮減することで，準最適な解を素早く学習する階層型強化学習手法を提案してきた [6, 7]．図 1 に提案手法の基本的な構成を示す．提案手法では，状態 $s \in S$ における行動 $a \in A$ の価値 $Q(s,a)$ と，複数の行動の組み合わせから構成されるスキル Λ の価値 $Q_\Lambda(s,\Lambda)$ を並列に学習する．S は状態集合，A は行動集合である．ここで，スキルは行動の順序集合 $a_\Lambda = \{a_1, a_2, \cdots, a_n\}$ 及びスキルを獲得した環境において a_Λ を実行した結果観測されるセンサ情報を抽象化した量 $M_\Lambda \in \mathbb{R}$ の直積 $\Lambda = (a_\Lambda, M_\Lambda)$ で定義される．各時刻において，行動はその状態における行動価値から確率的に決定されるが，このとき，行動価値には現在実行しているスキルの指定する行動に対して，スキル価値に基づいた一時的なバイアスが印可される．したがって，もし実行中のスキルが高い正の価値をもつものであった場合は，スキルの指定する行動が選択される確率は上昇する．逆にスキルの価値が負の値であれば，スキルの指定する行動が選択される確率は減少する．スキルによって行動選択に偏りが生じ，その結果，ロボットの探索行動にはスキルによって特徴づけられる一定の方向性が与えられる．

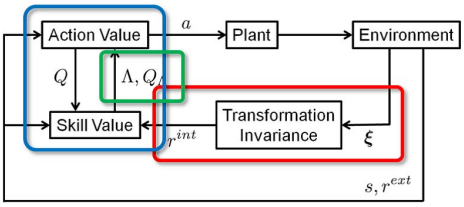
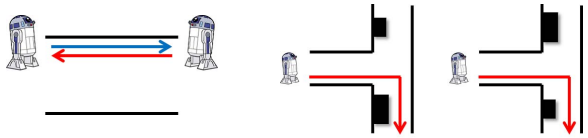


図 1 経験の再現性に基づく階層型強化学習の概要

行動価値の学習には一般的な TD 学習（Temporal Difference learning）を用いる．スキル価値の学習は SMDPs（Semi-Markov Decision Processes）において行われるが，タスクを記述する報酬信号に加えて，経験の再現性に対する内発的動機づけ [8] が与えられる．内発的報酬信号には，スキル獲得時に観測されたセンサ情報の系列に対する，現在のスキル実行の結果観測されるセンサ情報の系列の再現性を用いる．再現性の尺度としては，アファイン変換不変な特徴量 [9] を用いる．これにより，距離のような概念では測ることが難しい，図 2 のような経験の類似性を評価することが可能となる．図 2(a) では測距センサを搭載した移動ロボットが廊下において壁沿いに直進するスキルを実行する様子を表している．このとき，左から右へ移動する場合と，右から左へ移動する場合に観測されるセンサ入力は異なったものとなる．しかし，壁沿いに走行するというスキルの実行について考えると，進行方向に対してどちら側に壁が存在するかを区別せず，2 つ



(a) 廊下での壁沿い走行

(b) T 字路での右折

図 2 経験の同一性の尺度

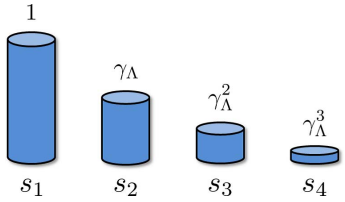


図 3 スキルにより訪問された状態スキル対に対する適格度トレース

の経験を同一視することができる．図 2(b) は，障害物の設置された T 字路において右折を行うスキルを実行する様子を表している．2 つの環境で障害物の大きさは異なるものの，右折スキルに関してこれらを区別する必要は小さいと考えられる．アファイン変換不変量を用いることで，上述のように不要な情報を捨象した，抽象的な経験の尺度を導入する．

アファイン変換不変量を経験の再現性に対する尺度とした内発的動機づけを与えることにより，ロボットの探索行動が特徴づけられる．すなわち，学習初期のほぼ完全に探索的な段階では，高い再現性を示す状態スキル対を手掛かりに，探索中心の候補となるパスを状態空間中に構成する．それらのパスのうち，正の報酬に到達したものが発見されると，さらに探索中心となるパスが限定される．その結果，スキルにしたがう指向性をもった探索戦略がとられ，状態行動空間全体を網羅的に探索することなく所望のタスクが達成されるため，学習速度が大幅に向上する．

3. スキル価値の学習における分配則

提案手法では，行動価値に加えて，ある状態 s において選択されたスキル Λ の価値である $Q_\Lambda(s,\Lambda)$ を学習する．スキルは行動の組み合わせから構成されるため，一般に行動よりもその実行に長い時間を要する．したがって，行動価値に比べてスキル価値が更新される頻度は低くなり，状態スキル空間においてスキル価値が更新される領域は相対的に限られたものとなる．加えて，提案手法ではスキル価値に基づいて行動選択過程にバイアスを与え，積極的な探索空間の縮減を図るため，ほとんどのスキル価値が更新されずに学習が行われる場合がある．

強化学習は本質的に試行錯誤的な探索を必要とする．そのため，提案手法にはこれまで探索的な行動選択により，スキル価値の未学習領域へ状態遷移することで学習が一時的に不安定化するという問題があった．本稿では，この問題を解決するため，スキル実行に伴う状態遷移の履歴から，内発的報酬信号を含む TD 誤差によって訪問した各状態におけるスキル価値を更新する適格度トレースを導入する．また，実際には訪問していない状態スキル対に対してスキル価値を汎化するため，スキル価値関数に対するフィルタリングを行う．

3.1 スキル実行の履歴に基づく適格度トレース

SMDPs における一般的な学習則では，ある状態 $s_1 \in S$ においてスキル Λ_j を実行した結果，遷移した状態 s_n における価値関数及び観測された報酬信号から，s_1 における Λ_j の価値 $Q_\Lambda(s_1,\Lambda_j)$ のみを更新する．しかし，ロボットの制御問題におけるスキルを想定した場合，類似の制御入力が開始から一定の時間は継続することが想定できる（移動ロボットの壁沿い走行，マニピュレータのリーチング動作等）．すなわち，スキルの開始状態近傍の局所的な領域について，空間的な距離に応じて TD 誤差を分配することには一定の合理性が見込まれる．そこで，本稿ではスキル価値の学習において，以下のような適格度トレースを用いる．

$s_1 \in S$ においてスキル Λ_j の実行を開始し，その後 $\{s_1, s_2, \cdots, s_i, \cdots, s_n\}$ と状態遷移したとする．ここで，n は Λ_j の実行に要する時間ステップである．状態 s におけるスキル Λ に対する適格度トレース $e_\Lambda(s,\Lambda)$ を以下のように更新する．

$$e_\Lambda(s,\Lambda) \leftarrow \begin{cases} \gamma_\Lambda^{i-1} & (s = s_i, \Lambda = \Lambda_j) \\ \lambda\gamma^{n-1} e_\Lambda(s,\Lambda_j) & (\text{otherwise}) \end{cases} \quad (1)$$

$\gamma_\Lambda \in [0,1]$ は，スキル実行に伴って訪問した状態群に与えるトレースの減衰を制御するパラメータである．$\lambda \in [0,1]$ 及び $\gamma \in [0,1]$ はそれぞれトレース減衰パラメータ及び割引率である．

式 (1) を適用することで，スキル実行に伴って訪問された状態群に対して，スキル実行開始からの時間経過によって指数関数的に減衰する重みで，スキル価値が一斉に更新される．図 3 にスキル実行によって訪問された状態群に対して割り当てられるトレースの概念図を示す．

3.2 バイラテラルフィルタによるスキル価値関数の汎化

式 (1) によって，スキル選択が行われた状態以外にも，スキル実行によって訪問された状態群へスキル価値を更新する領域を拡大する．この時点では，学習されるのは実際に訪問された状態群におけるスキル価値のみであるため，提案手法において構成される探索中心の候補となるパスを外れた多くの状態においては，スキル価値は更新されない．そこで，更新されたスキル価値関数に対するフィルタリング処理を施し，局所的な領域での学習結果をより広い領域へ分配する．

分配規則には各状態の空間的距離及びスキル価値の差を用いる．空間的な距離の近さは式 (1) と同様に，スキル価値の空間的連続性に着目するものである．しかし，1 つのスキルによって訪問される一連の状態群には，そのスキルに関して一定の連続的な性質を期待できる一方で，訪問されていない状態スキル対に関しては，非連続的な価値関数のエッジが存在する場合があ

る．そのため，ガウシアンフィルタ等の等方的な空間フィルタリングを行うと，スキル価値関数を特徴づける重要な変化を平滑化し，学習を阻害する場合があると考えられる．したがって，注目している状態におけるスキル価値と，その近傍の状態におけるスキル価値の差を考慮し，スキル価値関数のエッジを保存するフィルタリングを適用する．

以上の考えに基づいて，本稿ではバイラテラルフィルタ [10] をスキル価値関数に適用する．ここで，$\|\cdot\|$ は状態間のユークリッド距離を表すものとする．

$$G_\sigma(x) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2}{2\sigma^2}\right) \tag{2}$$

式 (2) を用いると，状態 $p \in S$ におけるスキル価値に対するフィルタリング処理は以下のようになる．

$$Q_\Lambda(p,\Lambda_j) \leftarrow \frac{1}{W_p}\sum_{q\in\bar{S}} Q_\Lambda(q,\Lambda_j)G_{\sigma_s}(\|p-q\|) \times G_{\sigma_r}(|Q_\Lambda(p,\Lambda_j)-Q_\Lambda(q,\Lambda_j)|) \tag{3}$$

$$W_p = \sum_{q\in\bar{S}} G_{\sigma_s}(\|p-q\|)G_{\sigma_r}(|Q_\Lambda(p,\Lambda_j)-Q_\Lambda(q,\Lambda_j)|) \tag{4}$$

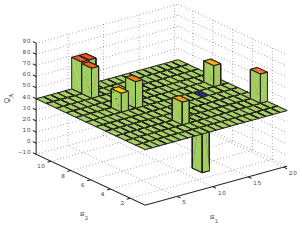
ここで，$\sigma_s \in \mathbb{R}_+$，$\sigma_r \in \mathbb{R}_+$ はそれぞれ空間的距離及びスキル価値の差に関するフィルタリングを調整するパラメータである．$\bar{S} \subset S$ はフィルタリングを適用する p を中心とした局所領域である．

実際に訪問していない状態スキル対に対するスキル価値の分配は，スキル利用に関する仮説形成のプロセスと捉えることができる．未学習領域を減少させることにより，多くの状態においてスキル価値に基づく行動選択へのバイアスが調整される．その結果，提案手法のもつ探索空間の縮減効果がより強く作用し，学習の収束が高速化すると考えられる．

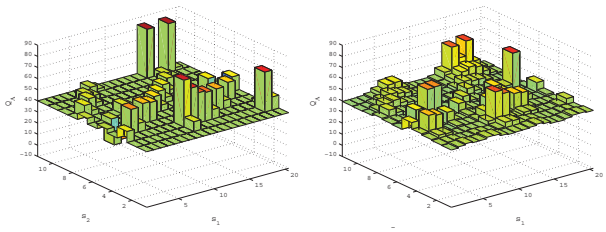
図 4 に，入替え更新トレース [1]（図 4(a)），スキル実行の履歴に基づく適格度トレース（図 4(b)），履歴に基づくトレースに加えてフィルタリング（図 4(c)）を適用して学習を行った場合に得られるスキル価値関数の例を示す．タスクは 4. 節で後述するマウンテンカー問題であり，それぞれ 3 エピソード目に学習されたスキル価値を表している．図 4(a) では多くの状態におけるスキル価値が初期値から変化していないが，図 4(b) ではスキルによって遷移した一連の状態に沿ってスキル価値が更新されている．さらに図 4(c) では，顕著にスキル価値の高い状態を保存しつつ，空間的に近い領域に価値が分配されていることがみてとれる．

4. シミュレーション実験

提案したスキル価値関数の汎化の学習効果を検証するため，シミュレーション実験を行った．タスクにはマウンテンカー問題 [11] を設定した．マウンテンカー問題の概念図を図 5 に示す．車は図中の谷底の初期状態から学習を開始し，右方向，左方向への移動及び停止の 3 つの行動を選択することができる．右側の山の頂上へ到達することが目的となる．しかし，動力不足のため，右方向への移動を選択し続けても目標状態へ到達することはできない．そのため，一度左方向へ山



(a) 入替え更新トレース



(b) 履歴に基づくトレース　　(c) フィルタリングあり

図 4 経験の再現性に基づく階層型強化学習の概要

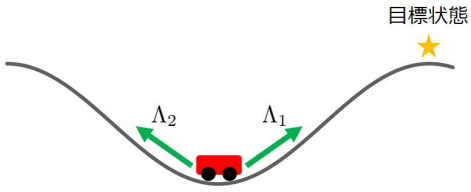


図 5 マウンテンカー問題

を登った後に，重力を利用して加速し右側の頂上へ移動する必要がある．

4.1 実験設定

スキルは初期状態から右方向及び左方向への移動を 40 ステップ継続することで得られる 2 つを利用した（図 5）．行動価値の学習には Sarsa(λ) [1] を，スキル価値の学習には内発的報酬信号を加えた SMDPs における Sarsa [6] を用いた．行動価値及びスキル価値の初期値は全ての状態について，それぞれ 0 及び 40 とした．行動選択及びスキル選択には温度定数 1 の soft-max 行動選択を用いた．報酬は目標状態において 100，1 ステップ経過毎に -1 とした．内発的報酬は $r_{int} = 80\exp(-dM^2/250)$ とした．ここで dM はスキルのもつ変換不変量とスキル実行の結果観測される変換不変量の差である．学習率 $\alpha = 0.5$，$\gamma = 1$，$\lambda = 0.95$，$\gamma_\Lambda = 0.9$，$\sigma_s = 0.5$，$\sigma_r = 40$ とした．行動選択時のバイアスの強度パラメータは $\beta = 0.5$ とした [6, 7]．エピソードの終了条件は目標状態への到達，または 1000 ステップの経過とした．

4.2 実験結果

Sarsa(λ)（図 6 中，黒のプロット），スキル価値の学習に入替え更新トレースを用いた提案手法 [6]（図 6 中，赤のプロット，以後従来手法と呼ぶ），スキル実行

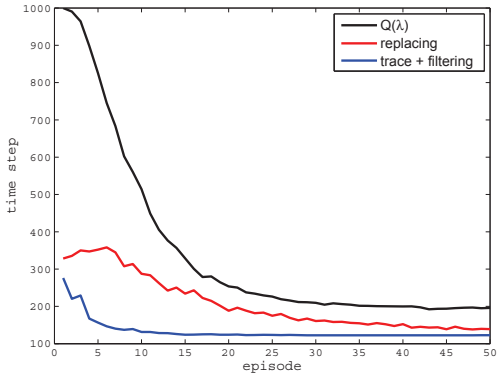


図 6 エピソード毎の経過時間

の履歴に基づくトレース及びフィルタ処理を施した提案手法（図 6 中，青のプロット，以後提案手法と呼ぶ）の 3 手法について，50 エピソードの実験を 300 回行った結果の平均を示す．図 6 の横軸はエピソード，縦軸は各エピソードの終了条件が満足されるまでに経過した時間ステップを表している．Sarsa(λ) と従来手法を比較すると，スキルを用いて経験の再現性に基づいた積極的な探索空間の縮減を行うことで，学習時間が短縮されることがわかる．しかし，スキル価値の更新頻度が低いため，従来手法では学習曲線は単調減少とはならず，一時的な学習の不安定化が生じている．それに対して，提案手法ではほぼ単調に目標状態へ到達するまでの時間が減少し，15 エピソード程度で学習が収束している．このことから，スキル実行によって訪問される局所領域に対する適格度トレースの導入と，バイラテラルフィルタによるスキル価値のエッジを保存したフィルタリングは，スキル実行の結果を適切に汎化することが可能であると考えられる．

4.3 考察

連続系を扱う強化学習の多くは，価値関数を関数近似手法を用いて構成するが，これは実際に訪問された状態の近傍に対する学習内容の汎化に相当する．したがって，本稿で提案した処理と同様に，空間的な距離に応じて学習内容を汎化することは可能であるという仮説に立脚したものと捉えられる．しかし，提案手法には空間的な近さだけでなく，スキルの実行によって訪問される状態群に対してトレースを設定し，スキルを利用した汎化を試みるという点に特徴がある．スキルという一連の行動選択の分節を用いた上述の操作は，能動的な探索空間の分節を促進する．また，バイラテラルフィルタを用いることで，スキルによって分節される状態における，価値関数の急峻な変化を保存することが可能となる．

本稿では，ルックアップテーブル型の強化学習をベースに手法を構成したが，より複雑な制御対象，センサ，環境における学習を実現するため，連続系への拡張が必要となる．スキル実行の履歴に基づく適格度トレースは，一般的な連続系における適格度トレースの自然な拡張として実現可能であると考えている．フィルタに関しても実装に問題はないと考えられるが，高次元の状態及び行動空間を想定するとオンラインでの計算負荷が大きなものとなる可能性がある．この点に関しては，バイラテラルフィルタの計算を高速化する手法の高次元のケースへの拡張や，関数近似された価値関数を用いて処理する方法等を検討している．

5. 結言

階層型強化学習における，時定数の大きな上位階層での価値関数の更新頻度の低さに起因する学習の不安定化を解決する手法を提案した．複数の行動の組み合わせから構成されるスキルを実行した結果訪問される状態群に対して，スキル価値の更新を行う適格度トレースを導入した．また，学習されたスキル価値に対して，空間的距離に応じた汎化とスキル価値のエッジを保存するフィルタ処理の適用を提案した．

提案手法をマウンテンカー問題に適用した結果，安定して学習を行うことが可能であり，かつ学習時間の大幅な短縮が実現されることを示した．今後は，連続系への拡張及び自律的なスキル手法の構築を行う．

参考文献

[1] R.S. Sutton, A.G. Barto: “Reinforcement Learning: An Introduction”, Cambridge, MA, MIT Press, 1998.
[2] A.G. Barto, S. Mahadevan: “Recent Advances in Hierarchical Reinforcement Learning”, Discrete Event Dynamical Systems: Theory and Applications, vol. 13, pp.341–379, 2003.
[3] T.G. Dietterich: “Hierarchical Reinforcement Learning with the MAXQ Value Function Decomposition”, Journal of Artificial Intelligence, vol. 13, pp.227–303, 2000.
[4] C.M. Vigorito, A.G. Barto: “Intrinsically Motivated Hierarchical Skill Learning in Structured Environments”, IEEE Transactions on Autonomous Mental Development, vol. 2, no. 2, pp.132–143, 2010.
[5] R. Der, G. Martius: “From Motor Babbling to Purposive Actions: Emerging Self-exploration in a Dynamical Systems Approach to Early Robot Development”, From Animals to Animats 9, Lecture Notes in Computer Science, Springer Berlin Heidelberg, vol. 4095, pp.406–421, 2006.
[6] 増山岳人, 山下淳, 淺間一: “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”, 日本機械学会論文集（C 編）, vol. 79, no. 798, pp.289–303, 2013.
[7] G. Masuyama, A. Yamashita, H. Asama: “Selective Exploration Exploiting Skills in Hierarchical Reinforcement Learning Framework”, Proceedings of the 26th IEEE/RSJ International Conference on Intelligent Robots and Systems, 2013.（発表予定）
[8] A.G. Barto: “Intrinsic Motivation and Reinforcement Learning”, Intrinsically Motivated Learning in Natural and Artificial Systems, Springer Berlin Heidelberg, pp.17–47, 2013.
[9] Y. Qiao, M. Suzuki, N. Minematsu: “Affine Invariant Features and Their Application to Speech Recognition”, Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, pp.4629–4632, 2009.
[10] C. Tomasi, R. Manduchi: “Bilateral Filtering for Gray and Color Images”, Proceedings of the 6th IEEE International Conference on Computer Vision, pp.839–846, 1998.
[11] S.P. Singh, R.S. Sutton: “Reinforcement Learning with Replacing Eligibility Traces”, Machine Learning, vol. 22, no. 1–3, pp.123–158, 1996.