CF-002

# モデルベース強化学習における自動計画を用いた探索戦略
# Exploration Strategy for Model-based Reinforcement Learning with Automated Planning

速水 陽平†‡　Amiri Saeid‡　Chandan Kishan‡　Shiqi Zhang‡　高玉 圭樹†
Yohei Hayamizu　Keiki Takadama

## 1. はじめに

人の代わりにタスクを行うためのロボットがものづくり産業や宇宙開発において利用されている [1] [2]．さらに，昨今の社会状況により，人の代わりに宅配作業や巡回を行うロボットの需要が高まっている．

このようなロボットの実現に向け，ロボットに代表されるエージェントは不確実性を含む環境において自律的に方策を学習することが求められる．環境モデルが与えられている場合，エージェントはマルコフ決定過程に基づいて方策の計算が可能だが，実環境において環境モデルは未知であることが多く，正確な環境モデルの設計が難しい．また，正確に環境モデルを設計しても，環境の変化に対応できないという問題がある．

強化学習は，報酬を基に各状態の評価値を見積もり，試行錯誤を通して各状態における適切な方策を学習し，環境モデルが与えられていない場合でも方策を獲得することが可能である．しかし，最適な方策を獲得するには十分な探索をする必要があるため，状態空間が膨大な場合，試行錯誤の頻度も膨大になる．この問題に対して，強化学習と自動計画を組み合わせることでエージェントの試行錯誤の回数を削減する研究がなされている [3]．自動計画は人の与えた行動ルールに基づいて行動系列を計画するアルゴリズムである．環境モデルが未知の場合は，自動計画により求められる行動系列が正確である保証はないが，強化学習の限定的な探索による方策学習を可能とする．一方で，この従来手法ではエージェントに与えるタスクが変化すると各状態の評価値も変化するため，再学習が必要となる．宅配作業や巡回のタスクはロボットが頻繁に様々な地点を訪問する必要があり，毎回要求されたタスクの方策を試行錯誤によって再学習するコストは大きい．

そこで本研究では，強化学習と自動計画の組み合わせによる探索空間の削減手法に加え，環境モデルの学習をすることで，再学習のコストを削減する手法を提案する．提案手法に対する類似研究として，モデルベース強化学習と確率的モデルを利用した自動計画を組み合わせた手法が提案されている [4]．しかし，確率的モデルは人の設計に依るところが大きいため，初期設計コストが決定的モデルと比較して大きいという問題がある．本研究では，環境モデルを正確に設計するのではなく，学習する点に焦点を当てる．本稿では，変化を伴う環境における方策学習の効率化を目的とし，提案手法の適用したエージェントが，方策学習時に必要のない探索を抑えつつ，複数のタスクを順次に達成できることを示す．具体的には自動計画を利用した評価値の初期化方法と，自動計画により生成された行動系列を利用した評価値更新方法を構築し，その有効性をシミュレータ上の迷路問題で検証する．

†電気通信大学　University of Electro-Communications
‡ニューヨーク州立大学ビンガムトン校　SUNY Binghamton

## 2. 背景知識

本節では，本研究で仮定とするマルコフ決定性仮定(MDP)と，モデルベース強化学習，および自動計画法としての解集合プログラミングについて記す．

### 2.1 マルコフ決定過程

マルコフ決定過程(Markov Decision Processes: MDPs)を定義することでエージェントの方策を決定すること [5]が可能である [6]．MDP のモデル D は$D =< S, A, P, R, \gamma >$から成る．$S$は状態集合，$A$は行動集合，$P$は状態遷移確率関数，$R$は報酬関数である．$\gamma$は割引率であり，環境の情報が伝播する割合を決定する．$0 < \gamma < 1$である．モデル D において，状態$s \in S$での最適方策$\pi^*$はベルマン方程式を利用して次のように求められる．

$$Q^*(s,a) = R(s,a) + \gamma \sum_{s' \in S} P(s,a,s')\, V^*(s')$$

$$\pi^*(s) \leftarrow argmax_a Q^*(s,a)$$

ここで$V^*(s)$は状態$s$における状態価値，$Q(s,a)$は状態$s$において行動$a$を選択した際の状態行動価値関数である．実際，ベルマン方程式は価値反復法や方策反復法などを用いて求められる [6]．また，上記の式を利用して状態行動価値関数$Q(s,a)$が最大となるような行動をとる方策を greedy 方策と呼ぶ．本研究では環境モデルが未知である状態を，状態遷移確率$P$および報酬関数$R$が未知であることと仮定する．

### 2.2 モデルベース強化学習

MDP 環境において状態遷移確率と報酬関数が未知であるとき，強化学習が利用される．特に，モデルベース強化学習(MBRL)はエージェントが状態遷移確率と報酬関数を学習し，プランニングを行うことで方策を学習する．モデルベース強化学習を利用する利点は 2 つある．1 つ目は、学習した MDP モデルを他のタスクに利用できるため，効率的な方策学習が可能である．2 つ目は，モデルベース強化学習のプランニングによる方策学習という特徴が人から与えられた環境に関する事前知識との組み合わせにより相乗効果が期待される点である．

### 2.3 解集合プログラミング

人から与えられた事前知識を用いたプランニングによりタスクを解決する方法として，ゴールまでの行動系列を導出する自動計画法がある．解集合プログラミング(Answer Set Programming: ASP)は，非単調推論の特徴を有した論理プログラミングの一つであり，不確定情報を含む常識推論(Commonsense Reasoning)が可能である [7]．行動系列を求める ASP は，自動計画法としても利用され，非単調性の観点から，考えられる全ての解を導出でき，行動系列の知識表現が容易であるという利点がある [5].

## 3. 指向型探索と方策学習

本節では，本研究の主題である強化学習と自動計画の組み合わせによる方策学習の効率化のため，指向型モデルベース Q 学習の提案を行う．強化学習と自動計画は MDP モデルを仮定したときに最適な行動系列を求めるという目的を同じくする一方で，それぞれが独立した領域として研究されてきた．指向型モデルベース Q 学習は，ASP よる行動系列の計画と，モデルベース強化学習による MDP モデルの学習を行い，指向的な探索によってタスクを解決するために必要な状態空間を優先的に探索することが可能である．指向型モデルベース Q 学習のフローチャートを図 1 に示す．まず，エージェントに対して事前知識を与え，自動計画によってタスクの達成に必要と思われる状態行動系列を計画する．得られた状態行動系列は Q 値の初期化(3.1 節)と仮想プランニングによる方策更新に利用される(3.2 節)．エージェントは事前知識を利用した方策更新と実際に行動して得られる観測を通して，試行錯誤的に方策を学習する．本研究では，事前知識を$\Pi(S,A,M)$と定義する．$M$はエージェントが達成するべきタスクであり，$M=(s_0,s_G)$とする．$s_0$はエージェントの初期状態であり，$s_G$は終端状態である．タスクが複数ある場合，$s_G$は状態集合$S_G \subset S$として表現されるが，簡単のため単一の状態とする．タスク$M$が与えられたとき，ASP による自動計画法は，行動数$L$の制約を考慮した上で，プラン$p$を一つ以上生成する．生成されるプランの集合を$\mathcal{H}$とすると，$p \in \mathcal{H}$は，以下の形で表現される．

$$p = \langle\langle s_0, a_0, s_1\rangle, \langle s_1, a_1, s_2\rangle, \ldots, \langle s_{G-1}, a_{G-1}, s_G\rangle\rangle$$

タスク$M$が与えられたときに，最小の行動数$l(\leq L)$となるプランを$p_l$とする．行動数の制約$L$はタスク$M$と同様に人が設計するパラメータであり，$L$が大きいほど，ASP は冗長なプランを計画する．本稿では，最短経路を求めるために ASP を利用するため，$L = 1.0$を用いる．最短経路が複数ある場合でも ASP は非単調性の特徴を持つため，複数の経路の集合$\mathcal{H}$を生成する．

### 3.1 自動計画を利用した楽観的初期化法

変化の伴う環境においては，ASP によって求められた最短経路は，事前知識$\Pi(S,A,M)$はしばしば環境モデルを簡略化するため，$\Pi(S,A,M)$について計画されたプランは楽観的なプランとして捉えることができる．例えば，エージェントが"$room\ R$へ行く"という行動をとることを考えると，”room R”の途中に障害物がある場合や環境が変化した場合はタスクの達成ができない可能性がある．楽観的初期化法の目的は，エージェントの探索空間を削減する事であり，ASP によって求められた楽観的なプランを利用して，状態行動価値関数$Q(s,a)$を初期化することで，エージェントがタスク$M$を達成する際に優先的に探索すると状態を決定する．状態行動価値関数$Q(s,a)$を初期化には，エージェントが得られる報酬の最大値$R_{max}$を用いる．計画した状態行動系列に誤りがあった場合は，その状態行動系列に対応する Q 値の値が減少する．

### 3.2 環境モデルを用いた仮想プランニング

楽観的初期化法により状態行動価値関数$Q(s,a)$が初期化されることで，エージェントは初期方策$\pi_0$を獲得する．この初期方策$\pi_0$を利用して，エージェントは環境との相互作用を通して最終的な最適方策を獲得する．本節では，指向型モデルベース強化学習の方策更新方法について述べる．具体的には，自動計画法によって計画される行動系列および，学習される環境モデルがどのように逐次的な方策更新に利用されているかを説明する．前節で自動計画法によって求められる行動系列は楽観的であると述べた．環境モデルを用いた仮想プランニングでは，楽観的なプランを利用して，仮想的な方策更新を反復して行う．仮想的な方策更新においては学習した環境モデルが必要となるが，環境モデルは，エージェントがある状態で実際に選択した行動によって観測されるデータに基づいて学習される．本研究では学習する環境モデルを，状態$s$で行動$a$をとって状態$s'$に遷移する時の状態遷移確率$P(s,a,s')$と，状態$s$で行動$a$をとる時に得られる報酬$R(s,a)$の予測モデルとする．状態$s$で行動$a$を選択した回数を$N(s,a)$，状態$s$で行動$a$を選択した際に状態$s'$に遷移した回数を$N(s,a,s')$として，状態遷移確率$P(s,a,s')$と報酬$R(s,a)$を以下のように定義する．

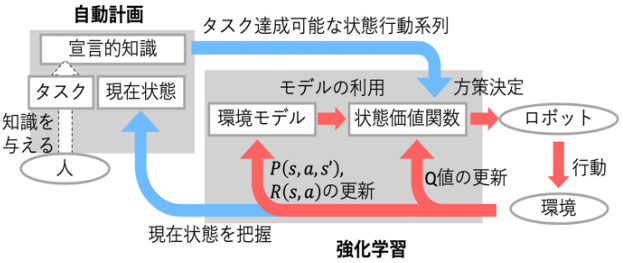


図 1：指向方モデルベース Q 学習のフローチャート

$$P(s,a,s') = \frac{N(s,a,s')}{N(s,a)}, R(s,a) = \frac{R(s,a)+r}{N(s,a)}$$

ここで，$r$は状態$s$で行動$a$を選択したときに実際に観測される報酬である．

## 4. 実験

指向型モデルベース強化学習が 1)探索空間を削減することで方策学習を効率化すること，2)学習した環境モデルを利用することで順次与えられるタスクに応じた方策を獲得することを，迷路問題におけるナビゲーションタスクを用いて検証する．

提案手法との比較手法として，モデルフリー強化学習手法である Q 学習，モデルベース強化学習手法である Dyna-Q [8]，自動計画とモデルフリー強化学習を組み合わせた手法である DARLING [3]を用いる．

### 4.1 ナビゲーションタスク

ナビゲーションタスクは，エージェントが初期状態から終端状態までの最適経路を求めるタスクである．本実験では図 2 に示すような$9 \times 9$マスの迷路問題を用いる．図 2 内の$S_0$はエージェントの初期状態，G1 および G2 は終端状態，太い黒線は壁，赤い太線はドアを表す．エージェントはまず，$M(S_0,G1)$が与えられる(Task1 とする)．次に，一定時間経過した後，$M(S_0,G2)$が与えられる(Task2 とする)．エージェントは Task1，Task2 をそれぞれ達成するために，状態$s \in S$に置いて，$[north, south, west, east, opendoor] \in A$を 1 つ選択する．各状態では一定の確率で選択した行動

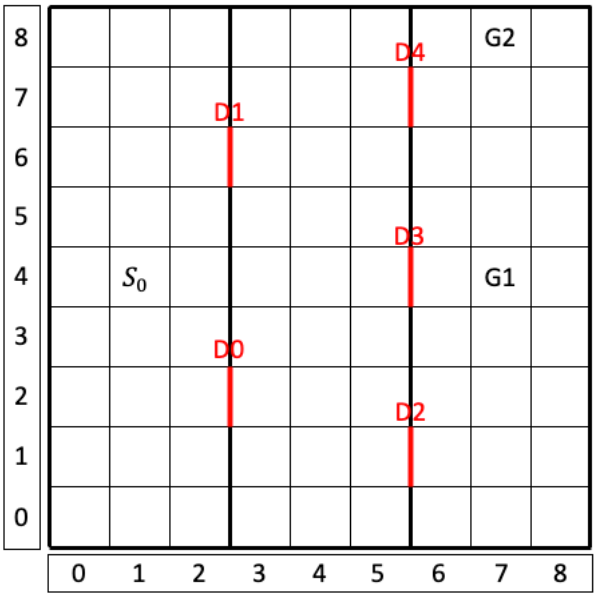


図 2：9×9マスの迷路問題

が失敗する．また，$opendoor$の行動に関しては，開閉確率が設けられており，ドアによってエージェントのドアを通り抜ける難易度が異なる．ナビゲーションタスクには 1 エピソードあたりのステップ数が設けられており，ステップ数の上限に達すると$S_0$から再スタートする．また，一定エピソードが経過した時点で，エージェントに与えられるタスクは Task1 から Task2 へ切り替わる．

ナビゲーションタスクにおいて，DARLING と提案手法に用いる環境モデルに関する事前知識の ASP による表現方法は次のとおりである．まず，環境の状態はエージェントの位置と壁・ドアの位置で与えられる．$pos(X,Y,I)$はステップ$I$においてエージェントの位置が$(X,Y)$にいることを示し，$obst(X,Y,D,I)$は方向$D$に壁があることを意味する．ASP における行動に関する知識は以下のように与えられる．

$$pos(X,Y+1,I+1):-north(I),pos(X,Y,I),I=0..n-1.$$

$pos(X,Y,I)$において$north(I)$を選択すると，状態が$pos(X,Y+1,I+1)$へ遷移することを表す．また，

$$:-north(I),pos(X,Y,I),obst(X,Y,no,I),I=0..n.$$

は方向$north$に$obst(X,Y,no,I)$がある場合の行動の制約を表している．

### 4.2 実験

各状態における状態遷移確率$P(s,a,s')$は0.95とし，1 つの行動につき，コスト$c=1$を必要とする．ドア$d0$から$d4$には開閉確率がそれぞれ設定されており，状態遷移確率$P(s,a,s')=0.95$に加えて，ドア$D$を通り抜ける確率$P(D)$はそれぞれ$P(d0)=0.8, P(d1)=1.0, P(d2)=0.95, P(d3)=0.6, P(d4)=0.95$とする．また，エージェントがそれぞれのタスクを達成した際に得られる報酬を$R_{max}=100$とする．さらに，Q 学習，Dyna-Q，DARLING，提案手法の各エージェントの方策は$\epsilon-greedy$法に従うこととし$\epsilon=0.9$とする．学習率$\alpha$および割引率$\gamma$は経験則から$\alpha=0.9, \gamma=0.9$に設定する．エージェントは 1 エピソードあたり 50 ステップの試行を行うことができ，2000 エピソード中 1000 エピソード経過した段階で，エージェントには Task1 から Task2 へのタスクの切り替えが行われるものとする．

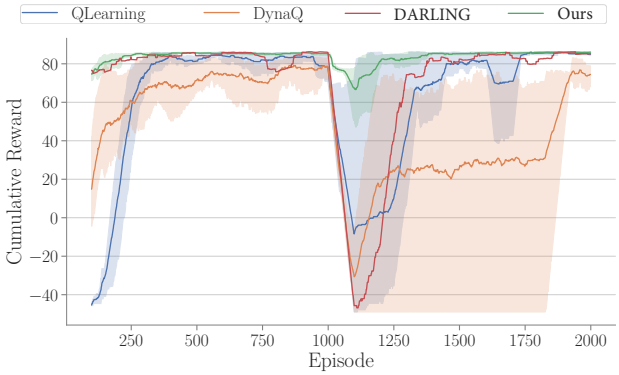


図 3：ナビゲーションタスクにおける累積報酬値の比較

図 3 は Task1 を学習した後に Task2 へ切り替えた時の，各手法の学習性能を比較した結果である．縦軸はエージェントが得た累積報酬値，横軸は学習回数を示し，色の違いは手法の違い，実線は累積報酬値の平均値，実線の周りの薄い色の領域は累積報酬値の分散値を示している．

### 4.3 考察

実験結果から，提案手法，DARLING はエピソードの早い段階で目的地に到達する方策を学習できており，誤った学習なしに(累積報酬値が負になることなしに)少ない試行回数で方策の学習(探索領域の削減)に成功しているが，提案手法は Task2 での学習性能の劣化(タスクが切り替わっても累積報酬が減少)がなく，これは，人が与えた宣言的知識によって該当する探索領域をロボットが優先的に探索しつつ(探索領域を絞り込みつつ)，環境モデルを学習することで早い方策の学習を実現していることを示している．

## 5. おわりに

本稿では，エージェントの方策学習の効率化を目的として，モデルベース強化学習と宣言的知識を利用した自動計画の相互補完的な特徴を利用した指向型モデルベース Q 学習の提案を行った． 迷路問題におけるナビゲーションタスクの実験を行うことで指向型モデルベース Q 学習の探索空間の削減度合いと類似タスクへの適応についての性能を評価，検証した．実験結果から，行動ルールなどの宣言的知識を利用しながら方策学習を行う指向型モデルベース Q 学習は従来手法よりも効率的な方策学習が可能であることが示された．

### 参考文献

[1] S. M. Goza, R. O. Ambrose, M. A. Diftler and I. M. Spain, "Telepresence control of the NASA/DARPA robonaut on a mobility platform," in Proceedings of the SIGCHI conference on Human factors in computing systems, 2004.
[2] Z. H. Khan, A. Siddique and C. W. Lee, "Robotics Utilization for Healthcare Digitization in Global COVID-19 Management," International Journal of Environmental Research and Public Health, vol. 17, no. 11, p. 3819, 2020.
[3] M. Leonetti, L. Iocchi and P. Stone, "A synthesis of automated planning and reinforcement learning for efficient, robust decision-making," Artif. Intell., vol. 241, pp. 103-130, 2016.
[4] J. H. A. Ng and R. P. Petrick, "Incremental learning of planning actions in model-based reinforcement learning," in Proceedings of the 28th International Joint Conference on Artificial Intelligence, 2019.
[5] V. Lifschitz, "Answer set programming and plan generation," Artificial Intelligence, vol. 138, pp. 39-54, 2002.

[6] M. L. Puterman, Markov Decision Processes.: Discrete Stochastic Dynamic Programming, John Wiley & Sons, 2014.
[7] E. Erdem, M. Gelfond and N. Leone, "Applications of answer set programming," AI Magazine, vol. 37, no. 3, pp. 53-68, 2016.
[8] R. S. Sutton, A. G. Barto, Reinforcement learning: An introduction, 2018.
[9] P. Khandelwal, S. Zhang, J. Sinapov, M. Leonetti, J. Thomason, F. Yang, I. Gori, M. Svetlik, P. Khante, V. Lifschitz and Others, "Bwibots: A platform for bridging the gap between ai and human--robot interaction research," The International Journal of Robotics Research, vol. 36, pp. 635-659, 2017.