4C2-3

ロバスト方策を用いた探索木によるベイジアン強化学習アプローチ

A Bayesian Reinforcement Learning Approach using Tree Search with Robust Policy

菱沼 徹　　泉田 啓
Toru Hishinuma　Kei Senda

京都大学大学院工学研究科航空宇宙工学専攻
Department of Aeronautics and Astronautics, Graduate School of Engineering, Kyoto University

This paper proposes "the tree search with robust leaves," a framework for an approximate modelbased Bayesian reinforcement learning problem for a real physical system. We focus on a situation that there is a critical state/action region in a real system and an agent knows the parametric structure of dynamics of the real system in advance. The proposed framework can learn an appropriate policy for a real system using only few real-world samples, by incorporating the tree search simulation and the pre-learning of a robust policy in simulation. It can also obtain safe and active searching behaviors in a real system. We show how to implement the framework in a real system. A mountain car task with a critical region is used for demonstrating the effectiveness of the proposed framework.

1. はじめに

強化学習は，未知な環境上で意思決定をするための有望な枠組みである [Sutton 98]．我々は，不確定性が存在する下での実システムの制御問題に対する強化学習の適用を考える．

ロボット強化学習の課題の一つは，"実サンプルの呪い" である [Kober 13]．実システムはシミュレーションとは異なり加速できないため，実サンプルの収集には時間がかかる．ロボットは摩耗や損耗により特性を変え，注意深いメンテナンスが必要になる．そのため，沢山の実サンプルを用いて収束するようなアルゴリズムは現実的ではない．また，ある領域の状態／行動をとると実システムに深刻なダメージを与えてしまうことがある．実システムはこの領域を探索してはならず，またこの領域を経由してタスクを達成する方策を選んではならない．

そこで，本研究では，安全な試行錯誤により少ない実サンプルを得て，不確定性が存在する実環境で動作する方策を得る学習方法を検討する．

ダイナミクスのパラメータ構造は既知であるが，実システムのパラメータ値は未知であるような状況を考える．具体的なパラメータ値は 1 つの実システムに固有であり，そのため別のシステムあるいはシミュレーションでは，知ることができない．さらに，実システムのことを知らずに探索方針や方策を決めてしまうと，エージェントは危険な領域に侵入してしまう可能性がある．危険を避けつつ能動的な探索挙動を実現するためには，先見情報（つまりパラメータ構造）の利用が必要である．

モデルベース・ベイジアン強化学習 [Duff 02] は，この状況で先見情報を統合する上で自然な枠組みである．この枠組みでは，実システムの不確定性は未知パラメータの事前分布として陽に記述される．全ての可能な試行錯誤過程の生起確率はベイズの法則に従って事前に知ることができる．原理的には，シミュレーションに基づく学習のみにより実システムに対するベイズ最適な挙動を得ることができる．そのため，先見情報と実サンプルの情報を最大限に利用することが可能であり，少ない回数の安全な探索による強化学習のために有望である．

この状況に対して，本研究では "ロバスト方策を葉とする探索木" という，モデルベース・ベイジアン強化学習の近似問題の枠組みを提案する．その主なアイデアは，有限の深さまでは探索木を用いて意思決定し，それ以降はロバスト方策に従うことである．また本研究では，ロバスト定常方策の事前学習と探索木を成長させるための UCT [Kocsis 06] を組み合わせて，提案する枠組みをエージェントが利用する方法を示す．

連絡先: 菱沼徹，　hishinuma.toru.43n@st.kyoto-u.ac.jp

2. モデルベース・ベイジアン強化学習

モデルベース・ベイジアン強化学習は，未知のマルコフ決定過程 (MDP) 上での意思決定問題として定式化される [Duff 02].

2.1 マルコフ決定過程

本研究では割引無し問題を扱う．MDP は，$(\mathcal{S},\mathcal{U},p,g)$ の 4 要素により定義される．ここで，$\mathcal{S}$ は状態の集合，$\mathcal{U}$ は行動の集合，$p:\mathcal{S}\times\mathcal{U}\times\mathcal{S}\to\mathbb{R}$ は状態遷移確率関数，$g:\mathcal{S}\times\mathcal{U}\times\mathcal{S}\to\mathbb{R}$ は 1 ステップコスト関数である．エージェントが現在の状態 s で行動 u を選択する場合，次の時刻で確率 $p(s'|s,u)$ で状態 s' へと遷移してコスト $g(s,u,s')$ を生じる．また，終端状態 s_0 は，$p(s_0|s_0,u)=1$，$g(s_0,u,s_0)=0$ を満たす．

方策 π はエージェントが行動選択を行うルールである．特に，定常方策は現時刻 t に対して，現在の状態 s^t のみに依存して現在の行動 u^t を選択し，履歴や時刻は陽に考慮されない．確定的定常方策は状態から行動への写像である．

コストが割引なしで蓄積する確率的最短経路問題を扱う [Bertsekas 96]．方策 π の性能の評価指標として，初期時刻の状態行動対 (s,u) に対する Q 値は，次式で定義される．

$$Q^\pi(s,u)\equiv E\left[\sum_{t=0}^{\infty} g(s^t,u^t,s^{t+1})\,\middle|\, s^0=s,u^0=u,\pi\right]$$

また，もう一つの評価指標である J 値は，次式で定義される．

$$J^\pi(s)\equiv E\left[Q^\pi(s,u)\,\middle|\, s^0=s,\pi\right] \tag{1}$$

$J^\pi(s)<\infty,\ \forall s$ であるような定常方策 π は，"プロパーである" と言われる [Bertsekas 96]．また，プロパーでない方策はインプロパー方策と呼ばれ，少なくとも 1 つの状態の J 値が無限大になり，エージェントは将来に少なくとも 1 つの状態から終端状態 s_0 へと到達することができない．

エージェントにとって有害な状態／行動という概念は，非常に大きな 1 ステップコストにより記述される．できるだけ小

さな J 値や Q 値を持つために，エージェントはこれらの状態／行動に直面しないような方策を学習する傾向にある．

2.2 ベイズ適応的マルコフ決定過程

ベイズ適応的 MDP（BAMDP）は，MDP の拡張として定義される [Duff 02]．$\theta \in \Theta$ を，MDP を指定する未知の時不変パラメータベクトルとする．本研究では，状態遷移確率のみが θ により変化し，$p(s'|s,u,\theta)$ と記述される場合を扱う．

エージェントは，未知パラメータ θ を直接観測できないが，モデル $p(s'|s,u,\theta)$（つまりパラメータから遷移確率への写像）と未知な θ の事前分布 $b^0(\cdot)$ を前以て持つ．遷移 (s,u,s') が観測されたとき，事後分布 b^t は次のベイズの法則に従う．

$$b^{t+1}(\theta) = \frac{p(s'|s,u,\theta)b^t(\theta)}{\int_{\theta'} p(s'|s,u,\theta')b^t(\theta')d\theta'} \tag{2}$$

BAMDP に対する J 値は次のように定義される．

$$\bar{J}^\pi([s,b^0]) \equiv E\left[J_\theta^\pi(s) \,\middle|\, b^0\right]$$

ここで，$J_\theta^\pi(s)$ は，MDP パラメータ θ に対する J 値である．

履歴 $h^t \equiv [b^0, s^0, u^0, \cdots, s^{t-1}, u^{t-1}, s^t]$ に対して最適方策 $\pi^t : h^t \mapsto u$ を導く問題と，未知パラメータの事後確率分布 $b(\theta) \equiv p(\theta|h)$ として最適方策 $\pi : [s,b] \mapsto u$ を導く問題は等価である．Bellman 方程式は以下のように書ける．

$$\bar{J}^*([s,b]) = \min_u E\left[g(s,u,s') + \bar{J}^*([s',b']) \,\middle|\, s,b,u\right] \tag{3}$$

ここで，次の時刻の信念状態 b' は，現在の信念 b と観測された遷移 (s,u,s') にベイズの法則 (2) を適用して得る．

3. ロバスト方策を用いる探索木

3.1 BAMDP に対するロバスト定常方策

MDP に対する最適方策は現在の状態 s^t のみに基づいて行動選択し，定常方策である．しかし，BAMDP に対する最適方策は $[s,b]$ に基づいて行動選択し，定常方策ではない [Senda 14]．定常方策は BAMDP に対して一般に最適ではないが，より少ない計算量により得られる準最適解としてしばしば役に立つ．

BAMDP に対する準最適な定常方策を求める問題は，確率的定常方策の集合を Π_s として，次のように書くことができる．

$$\bar{J}_s^*([s,b]) \equiv \min_{\pi\in\Pi_s} \bar{J}^\pi([s,b]) = \min_{\pi\in\Pi_s} E\left[J_\theta^\pi(s) \,\middle|\, b\right] \tag{4}$$

もし定常方策 π が $\bar{J}^\pi([s,b]) < \infty,\ \forall s$ を満たせば，π は Θ の中の全ての θ に対してプロパーである．そのため，BAMDP に対するプロパー方策は，どのような MDP パラメータ θ に対してもタスクを達成することができることを保証されている．

定常方策は，初期時刻以降のオンライン遷移情報により変更されず，そのため未知パラメータ θ を知ることを狙いとした探索的行動を選択しない．本研究では，どのような不確定性に対してもタスク達成が保証される固定された制御器という意味で，プロパー定常方策をロバスト方策と見なす．

3.2 緩和問題の概要

ロバスト方策を用いる前進探索木は，終端コストとして 3.1 節の定常方策の J 値を用いる finite horizon 問題である．固定 planning horizon を D とする．深さ $d = D$ において $[s,b]$ でタスクを終了する終端コストを，$\bar{J'}^*([s,b],D) \equiv \bar{J}_s^*([s,b])$ で与える．深さ $d < D$ の J 値は，次の Bellman 方程式に従う．

$$\bar{J'}^*([s,b],d) = \min_u E\left[g(s,u,s') + \bar{J'}^*([s',b'],d+1) \,\middle|\, s,b,u\right] \tag{5}$$

この緩和問題の解は，有限深さ D に到達する前には探索木による行動選択を行い，その後には深さ D 到達時の $[s,b]$ に対して (4) を最小化する定常方策 π_{sb} に従う．この問題は探索深さ D の範囲内でロバスト方策による挙動を含めて探索を比較しているため，その解はロバスト方策よりも良い評価指標を持ち，“安全かつ能動的な試行錯誤挙動” が得られる．

Algorithm 1 Q-Learning for BAMDP

```
function Pre-Q-learning(b)
initialize Q^{π_s}(s', u'), ∀s', u'
repeat
  initialize s
  θ ~ b^0(·)
  SimulateQagentEpisode(s, θ)
until reach number of times of episode
return Q^{π_s}

function SimulateQagentEpisode(s, θ)
repeat
  choose u from s using ε-policy derived from Q^{π_s}(s, ·)
  s' ~ p(·|s, u, θ)
  Q^{π_s}(s, u) ← (1 − γ)Q^{π_s}(s, u)
                  + γ {g(s, u, s') + min_{u'} Q^{π_s}(s', u')}
  s ← s'
until s = s_0
```

4. 緩和問題の実装方法

大規模タスクあるいは連続的タスクにおいては，提案する緩和問題 (5) を解くことでさえ計算量的に難しい．また，生成モデル（運動方程式等の連続時間ダイナミクス）は直接的に与えられる一方で，対応する離散時間ダイナミクス $p(s'|s,u,\theta)$ はすぐには利用できないことがしばしばある．本節では，このような状況下で我々の近似問題を実装するアプローチを述べる．

4.1 定常方策のオフライン事前学習

我々は，できるだけオンライン計算資源を木探索に投資するために，葉ノードに対して良い定常方策を与えるための計算を事前にオフラインで実行する．しかしながら，可能な $[s,b]$ の全てに対して最適化問題 (4) の解を事前にオフラインで用意することは不可能である．本研究では，一番単純な場合として，BAMDP に対する確定的プロパー定常方策を 1 つのみオフライン事前学習することを考える．Algorithm 1 は，Q 学習と信念状態からの MDP サンプリングを組み合わせた手法であり，γ は学習率である．

4.2 UCT 探索

現在の $[s,b]$ からの探索戦略を見出すために，シミュレーション上でモンテカルロ木探索を行う．[Guez 12] と同様に，有望な枝に計算資源を割り当てるために UCT [Kocsis 06] を用いる．探索木シミュレーションの行動選択において，探索ボーナス（負のコスト）$c\sqrt{\frac{\log N(s,b)}{N([s,b],u')}}$ を追加する．

Algorithm 2 UCT search simulation

function SimulateTree(h^d, θ, d)
if $d = D$ **then**
  return $\min_u Q^{\pi_s}(s^D, u)$
end if
if $N(h) = 0$ **then**
  for $u \in \mathcal{U}$ **do do**
    $N(h^d, u) \leftarrow 0$
    $Q_{tree}(h^d, u) \leftarrow 0$
  end for
  $u \leftarrow \arg\min_{u'} Q^{\pi_s}(s^D, u')$
  $s' \sim p(\cdot|s^d, u, \theta)$
  $q \leftarrow g(s^d, u, s') + \min_{u'} Q^{\pi_s}_{\theta}(s', u')$
  $N(h^d) \leftarrow 1$
  $N(h^d, u) \leftarrow 1$
  $Q_{tree}(h^d, u) \leftarrow q$
  return q
end if
$u \leftarrow \arg\min_{u'} \left\{ Q_{tree}(h^d, u') - c\sqrt{\frac{\log N(h^d)}{N(h^d, u')}} \right\}$
$s' \sim p(\cdot|s^d, u, \theta)$
$q \leftarrow g(s^d, u, s') + \text{SimulateTree}([h^d, u, s'], \theta, d+1)$
$N(h^d) \leftarrow N(h^d) + 1$
$N(h^d, u) \leftarrow N(h^d, u) + 1$
$Q_{tree}(h^d, u) \leftarrow Q_{tree}(h^d, u) + \frac{q - Q_{tree}(h^d, u)}{N(h^d, u)}$
return q

Algorithm 2 は，現在の $[s, b]$ と MDP のシミュレーションパラメータ θ が与えられた場合の探索木アルゴリズムである．N はカウンター関数，π_s は事前に用意した BAMDP に対するロバスト定常方策である．2.2 節で述べたように，h^d と $[s^d, b^d]$ は等価な意思決定のための情報を持つ．MDP パラメータ θ の生成モデルを用いて (s, u) から次の状態 s' をサンプルすることを $s' \sim p(\cdot|s, u, \theta)$ により表す．

Algorithm 2 では，葉ノードの値を与えるために，事前学習によるロバスト定常方策の Q 値を用い，[Guez 12] における探索木中でのベイズ事後分布の計算と葉ノードでの rollout を省略する．これは，プロパー定常方策はパラメータ θ の変動に対してロバストであるように学習されるため，その性能指標は rollout よりも保守的に見積もられる，という考えに基づく．

4.3 オンラインエージェントへの実装

4.1, 4.2 節の方法をオンラインエージェントに実装する手順を Algorithm 3 に示す．ここで，現在の状態 s と行動 u を条件として，実システム（パラメータ θ^* を伴う生成モデル）から次の状態 s' を得ることを $s' \sim p(\cdot|s, u, \theta^*)$ により表す．エージェントは θ^* を直接観測できないとする．

5. 数値実験

5.1 問題設定

独自の設定を追加した Mountain Car タスクを用いて，本研究のアプローチを検証する．エージェント（車）のゴールは，右の山頂にたどり着くことである．標準的な設定では，エンジンよりも重力の方が強いため，車はゴールに向かって加速するだけでは山頂にたどり着けない．この場合の最適方策による挙動は，まず左の山の中腹まで登り，その位置エネルギーを利用して加速して右の山頂へ到達する．

本研究では，さらに次の設定を追加する．(i) 左の山には崖

Algorithm 3 Implementation

[**Offline Phase**]
Input: prior distribution b
Q^{π_s} = Pre-Q-learning(b) (see Section 4.1)

[**Online Phase**]
Input: $t = 0$, initial state s, real MDP parameter θ^*
repeat
  $\theta \sim b(\cdot)$
  SimulateTree($h, \theta, 0$) (see Section 4.2)
until time out
repeat
  if $t < D$ **then**
    $u \leftarrow \arg\min_{u'} Q_{tree}(h, u')$
  else
    $u \leftarrow \arg\min_{u'} Q^{\pi_s}(s, u')$
  end if
  $s' \sim p(\cdot|s, u, \theta^*)$
  $h \leftarrow [h, u, s']$
  $s \leftarrow s'$
  $t \leftarrow t + 1$
until $s = s_0$

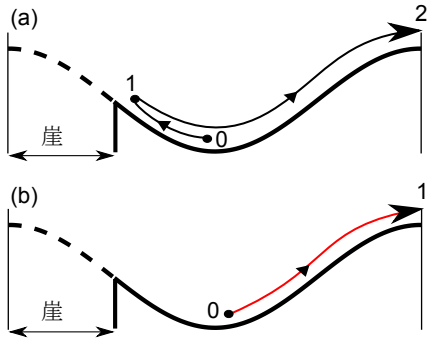

図 1: Mountain Car タスク．

があり，車はその領域に立ち入ってはならない．(ii) 車は追加コストを支払うことでブースターを利用することが可能で，これを用いれば左の山に登らずに右の山頂に到達できる．

既知の MDP に対する最適挙動は 2 種類存在する．図 1(a)：左への行動では崖から落ちないため，標準的な設定と同様に左の山を経由してゴールに向かう．図 1(b)：左への行動では崖から落ちるため，ブースターを利用して左の山に経由せずにゴールへ向かう．赤線は，ブースターの利用を表す．どちらの方策を取るべきかは，左への行動で崖から落ちる可能性に依る．

具体的には，次の生成モデルを考える．

$$\begin{cases} \dot{x} = v \\ \dot{v} = -g \sin x + (\theta + \sigma)u \end{cases} \tag{6}$$

状態変数は $s = [x, v]$ である．離散行動空間は，$\mathcal{U} = \{-1, 1, 3\}$ である．ここで，$u = \pm 1$ は標準的な設定のエンジンを，$u = 3$ はこの例題に特有のブースターを用いる事を意味する．

意思決定のための離散ステップを，t と表記する．微分方程式 (6) の中の連続時間を，t' により表記する．この例題では，時間幅 $\Delta t' = 1.2$ 毎に，車が行動を選択する．

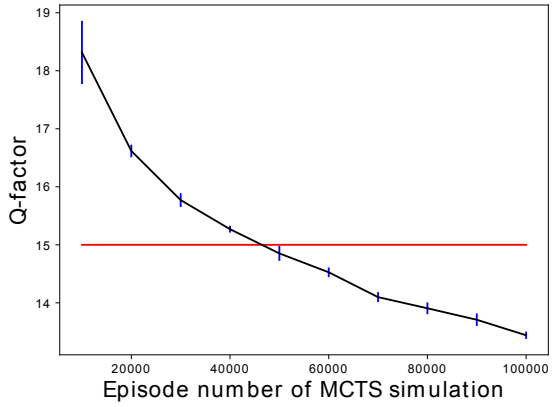

図 2: 初期ノードの Q 値（黒は $u = 1$，赤は $u = 3$）.

エンジン出力外乱 σ は，行動選択と同じタイミングで，区間 $[-0.05,\ 0.05]$ 上の連続一様分布からサンプルされる．車にとって既知であるのは，この外乱の特性，重力定数 $g = 9.8$，式 (6) の構造である．一方で，エンジンの出力平均 θ は車にとって未知なパラメータであり，以下の事前分布を考える．

$$b^0(\theta) = \begin{cases} 1 & (5.0 \leq \theta \leq 6.0) \\ 0 & (\text{otherwise}) \end{cases} \tag{7}$$

この問題の 1 ステップコストは，次式により定義される．

$$g([x, v], u, [x', v']) = \begin{cases} 300 & (x' < -0.44\pi) \\ 15 & (x' > -0.44\pi,\ u = 3) \\ 1 & (\text{otherwise}) \end{cases}$$

ここで，危険エリアは $[-\pi,\ -0.44\pi]$ であり，この領域への遷移には 300 の大きなコストでペナルティが課される．車は，実システム上では，この状態遷移をしてはならない．一方で，探索木とロバスト定常方策を計算するシミュレーション内では，この状態遷移を経験できる．

この例題では，意思決定の学習を簡単にするために，タイル空間を用いている．x 方向の区間 $[-\pi,\ \pi]$ が 75 個に分割され，また v 方向の区間 $[-8.0,\ 8.0]$ が同じく 75 に分割される．ここで，崖の位置 $x = -0.44\pi$ は，タイルの境界と一致する．

MDP のパラメータが $\theta^* = 5.0$ の場合には，図 1(a) の最適挙動をとる．逆に，$\theta^* = 6.0$ の場合には，図 1(b) の最適挙動をとる．実システムのパラメータが未知である場合の BAMDP に対するロバスト定常方策は，図 1(b) の挙動をとる．

5.2 結果

初期状態を $s^0 = \mathbf{0}$ で固定する．Algorithm 3 において探索シミュレーションのエピソード回数を変えた時の，行動 $u = 1,\ 3$ に対する初期ノード $[s^0, b^0]$ の Q 値の計算結果を，図 2 に示す（行動 $u = -1$ の Q 値は 50 以上である）．各エピソード回数につき 10 回ずつ数値実験を行った．なお，Algorithm 2 中で，深さ $D = 3$ と UCT 探索定数 $c = 200$ とした．

エピソード回数を十分に取ると，初期ノードから行動 $u = 1$ を選択した．それ以降の挙動は，実システムのパラメータが $\theta^* = 5.0$ の場合には図 3(c)，$\theta^* = 6.0$ の場合には図 3(d) となった．初期時刻における行動 $u = 1$ は，図 1 には見られず，

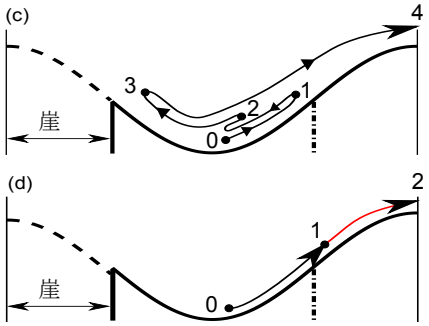

図 3: 提案手法から得られる挙動.

安全に山の登りやすさを確認するという意味のみを持つ．もし崖から落ちる可能性がないなら，図 3(c) の挙動は，最初からブースターを使うよりも少ないコストで済む．また，もし崖から落ちる可能性があるとしても，図 3(d) の挙動は最初からブースターを使うよりも，わずかに劣るのみである．このように，安全かつ能動的な探索挙動が，提案手法から得られた．

6. おわりに

本研究では“ロバスト方策を葉とする探索木 ”という，モデルベース・ベイジアン強化学習の近似問題の枠組みとその実装アプローチをを提案した．提案する枠組みを実装する方法を示し，また Mountain Car タスクを通じてその有効性を検証した．将来の課題は，現実のより大規模な物理システムに適用することにより拡張性を評価する事である．

参考文献

[Sutton 98] R. S. Sutton and A. G. Barto: *Reinforcement Learning: An Introduction*, MIT Press (1998).

[Kober 13] J. Kober, J. Bagnell, J. Andrew, and J. Peters: Reinforcement Learning in robotics: A SurveyThe *International Journal of Robotics Research* (2013).

[Duff 02] M. O. Duff: *Optimal Learning: Computational procedures for Bayes-adaptive Markov Decision Processes*, PhD thesis, University of Massachusetts Amherst (2002).

[Kocsis 06] L. Kocsis and C. Szepesvari: Bandit based monte-carlo planning, In *European conference on machine learning*, pp. 282-293 (2006).

[Bertsekas 96] D. P. Bertsekas and J. N. Tsitsiklis: *Neuro-Dynamic Programming*, Athena Scientific (1996).

[Senda 14] K. Senda amd Y. Tani: Autonomous robust skill generation using reinforcement learning with plant variation, *Advances in Mechanical Engineering* (2014).

[Guez 12] A. Guez, D. Silver, and P. Dayan: Efficient Bayes-adaptive reinforcement learning using sample-based search, In *Advances in Neural Information Processing Systems*, pp.1025-1033 (2012).