モデルの不確実性を考慮した決定理論的計画立案

末松伸朗† 林 朗†

アクションの効果が非決定的な問題領域での計画立案について，マルコフ決定過程理論に基づく決定理論的計画立案が提案されている．本論文では，これまでの決定理論的計画立案研究の中で考慮されなかったモデルの不確実性を考慮した計画立案方法を提案する．モデルの不確実性は（1）強化学習において，探検（モデル学習）途中である場合や，(2) マルチエージェント環境で，エージェントが異なる戦略をもった複数の匿名対戦者と繰り返し対戦する場合などに現れる．本論文では，候補モデルの有限集合とその集合上の確率分布（確信度）が与えられると仮定する．すると，最適な政策とは平均報酬（ステップあたりの報酬の平均値）のモデルの不確実性に関する期待値を最大化する政策として定義できる．筆者らは，確率的政策の空間において平均報酬の期待値の極大値を求めるための政策反復アルゴリズムを開発した．

Consideration of Model Uncertainty in Decision Theoretic Planning

NOBUO SUEMATSU† and AKIRA HAYASHI†

The theory of Markov decision processes provides a basis for planning in stochastic domains where the effects of actions are not deterministic. In this paper, we consider another dimension of uncertainty, *model uncertainty*. The uncertainty on the model naturally arises (1) when exploration is not complete in model learning, or (2) when the agent is supposed to repeatedly interact with multiple anonymous opponents who have different strategies. We assume that a set of candidate models and the probability distribution over the set are given. Then the optimal policy will be one which maximizes *the expectation* of the average reward per step (gain) with respect to the model uncertainty. We have developed a policy iteration algorithm to find a local maximum of the expected gain in the space of stochastic policies.

1. はじめに

アクションの効果が決定的ではなかったり，初期条件が確実には知られていなかったり，互いに競合するいくつかの目的をトレードオフして最適プランを見つける必要がある問題領域での計画立案について，マルコフ決定過程理論[6),11)]に基づく決定理論的計画立案が提案されている[3),4)]．本論文では，モデルの不確実性という，これまでの決定理論的計画立案研究の中で考慮されなかった不確実性を考慮した計画立案方法を提案する．

モデルの不確実性は，モデル学習において探検が完了していないときに現れる．また，マルチエージェント環境では，モデルの不確実性がまったく異なる意味で現れる．有名な繰返し囚人のジレンマのトーナメント問題[1)]では，個々のエージェントは異なる戦略を持った複数の匿名対戦者と繰り返し対戦する設定になっている．もし，ある特定の対戦者との対戦の繰返しを1つのモデルとしてモデル化すれば，複数の対戦者との対戦は，複数のモデルとその上の確率分布としてモデル化できる．いずれの場合も，モデルの不確実性を無視して，1つの候補モデルにコミットして計画立案すれば，良くない結果を招く危険性がある．

筆者らは，モデルの不確実性の度合いに関する推定値を利用するベイズ的アプローチをとる．本論文では，候補モデルの有限集合とその集合上の確率分布（確信度）が与えられると仮定する．すると，最適な政策とは平均報酬（ステップあたりの報酬の平均値）のモデルの不確実性に対する期待値を最大化する政策として定義できる．筆者らは，平均報酬の期待値を局所最大化するように，確率的政策を改良するための，政策反復アルゴリズムを開発した．このアルゴリズムの骨子はモデルの不確実性を，状態の不確実性に読み替える点にある．たとえ，そのぞれのモデル候補について，状態が完全観測である（MDP 仮定）としても，エージェントはモデル候補のうちどの候補が真のモデルであるかは分からない．このようにして，モデルの不確

† 広島市立大学情報科学部
Faculty of Information Sciences, Hiroshima City University

実性問題は，部分観測マルコフ過程（Partially Observable Markovian Decision Processes, POMDPs）において研究されている知覚エイリアシング問題[7),14)]の一例と見なすことが可能になる．

モデルの不確実性を考慮した計画立案方法に関する研究は少ない．Schneider[13)]は，倒立振子をバランスさせるコントローラが与えられたとき，けっして棒を倒すことなく台車を移動する問題を強化学習問題として考えた．学習初期にはモデルに関する知識が不確実であるにもかかわらず，棒を倒してはならない点が難しい．彼は，モデルの不確実性を取り入れた値反復計算を行い，用心深いダイナミック・プログラミングと呼んだ．しかし，得られる政策の最適性の検討はなされておらず，また，モデルの不確実性が学習につれ急速に減少する場合を想定しているようであり，一般化は難しいと考えられる．

Ben-Porath[2)]は，繰返しのある2人ゲームにおいて，決定的オートマトンの集合（候補オートマトン）とその上の確率分布で表される相手の混合戦略に対し，平均報酬の期待値を最大化する決定的オートマトンを求める問題を研究した．彼は最良の決定的オートマトンを相手の候補オートマトンの積オートマトンの大きさの多項式時間で見つけるアルゴリズムを示した．候補の数が増加するにつれて，積オートマトンは急激に大きくなるので，このアルゴリズムの計算量的負荷は大きい．決定的オートマトンは内部状態を表現できるという長所があるが，非決定的な遷移を表現できない点で本論文でモデル化の対象とするマルコフ決定過程と大きく異なる．

2. 問題の定式化

2.1 不確実なモデル問題

エージェントのセンサー入力を $\{s \mid s \in S\}$，アクションを $\{a \mid a \in A\}$ とする．$\pi(a \mid s)$ をエージェントの確率的政策とする．確率的政策はセンサー入力の集合 S から，アクションの集合 A の上の確率分布（PDs）への写像である☆．

$\hat{M} = \{M_i \mid 1 \leq i \leq n\}$ を，センサー入力の集合 S とアクションの集合 A を共有するMDPモデルの有限集合とする．各モデル M_i の状態は，センサー入力 s から識別可能であると仮定し，s を各モデル M_i の状態を表すのにも用いることにする．各モデル M_i に対して，状態遷移確率 $\{P_i(s,a,s') \mid s \in S, a \in A, s' \in S\}$ と直接報酬 $\{R_i(s,a,s') \mid s \in S, a \in A, s' \in S\}$ が定義されているとする．ここで，$P_i(s,a,s')$ は，モデル M_i において状態 s にあるとき，行動 a によって状態 s' へ遷移する確率であり，$R_i(s,a,s')$ は，そのような遷移が起きたときエージェントが受け取る直接報酬である．また，各モデル M_i は任意の政策に関して，エルゴード的であると仮定する．

政策 π が与えられれば，各モデル M_i に対して，その政策のもとでの状態 s の状態占有確率 $P_i^\pi(s)$，平均報酬 R_i^π を以下の式を用いて計算することができる[6)]．

$$P_i^\pi(s) = \sum_{s' \in S} P_i^\pi(s') \times \sum_{a \in A} \pi(a \mid s') P_i(s', a, s) \tag{1}$$

$$\sum_{s \in S} P_i^\pi(s) = 1 \tag{2}$$

$$R_i^\pi = \sum_{s \in S} P_i^\pi(s) \sum_{a \in A} \pi(a \mid s) \times \sum_{s' \in S} P_i(s,a,s') R_i(s,a,s') \tag{3}$$

なお，エルゴード的な有限マルコフ過程の状態占有確率に関する議論，たとえば教科書9)から，任意の s について $P_i^\pi(s) > 0$ であることを容易に示すことができる．

$\hat{q} = (q_1, q_2, \ldots, q_n)$，$\sum_i q_i = 1$ をモデルの集合 $\hat{M}$ 上の確率分布とし，政策 π の $(\hat{M}, \hat{q})$ に対する平均報酬の期待値 $\hat{R}^\pi$ は次の式で計算できる．

$$\hat{R}^\pi = \sum_{i=1}^{n} q_i R_i^\pi \tag{4}$$

ここで，筆者らが解決しようとしている問題を定義する準備ができた．

不確実なモデル問題 $(S, A, \hat{M}, \hat{q})$ が与えられたとき，$\hat{R}^\pi$ を最大化する政策 π を求めよ．

2.2 例　　題

図1にある2つのMDPモデル候補について考える．この例では，$\hat{M} = \{M_1, M_2\}$，かつ $\hat{q} = (1/2, 1/2)$ であるとする．政策 π を以下のように定める．$\pi(a|s_1) = \pi_1^a$，$\pi(b|s_1) = 1 - \pi_1^a$，$\pi(a|s_2) = \pi_2^a$，$\pi(b|s_2) = 1 - \pi_2^a$．すると，定義より

$$P_1^\pi(s_1) = \frac{\pi_2^a + \Delta - 2\pi_2^a\Delta}{1 - \pi_1^a + 2\pi_1^a\Delta + \pi_2^a - 2\pi_2^a\Delta}$$

$$P_1^\pi(s_2) = \frac{1 - \pi_1^a - \Delta + 2\pi_1^a\Delta}{1 - \pi_1^a + 2\pi_1^a\Delta + \pi_2^a - 2\pi_2^a\Delta}$$

☆ 本論文では，定常政策のみを扱う．過渡政策（適応政策）は考えない．

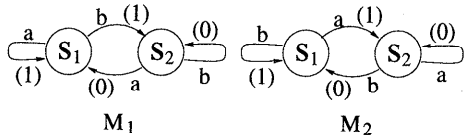


図1 2つのMDPモデル候補．括弧の中の数字は直接報酬を示す．それぞれのモデルをエルゴード的にするために，Δ の確率で，微少なアクション・エラーが起きるようにした．したがって，$P_1(s_1, a, s_1) = 1 - \Delta$，$P_1(s_1, a, s_2) = \Delta$，*etc.*

Fig. 1 Two candidate MDP models. The numbers in the parentheses are immediate rewards. We put Δ action error to make the models ergodic. Therefore $P_1(s_1, a, s_1) = 1 - \Delta$, $P_1(s_1, a, s_2) = \Delta$, *etc.*

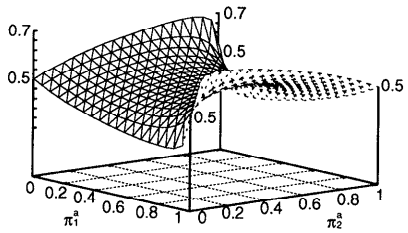


図2 $\Delta = 0.01$ のときの $\hat{R}^\pi(\pi_1^a, \pi_2^a)$

Fig. 2 $\hat{R}^\pi(\pi_1^a, \pi_2^a)$ for $\Delta = 0.01$.

$$R_1^\pi = \frac{\pi_2^a + \Delta - 2\pi_2^a\Delta}{1 - \pi_1^a + 2\pi_1^a\Delta + \pi_2^a - 2\pi_2^a\Delta}$$

同様に，M_2 について

$$R_2^\pi = \frac{1 - \Delta - \pi_2^a + 2\pi_2^a\Delta}{1 + \pi_1^a - 2\pi_1^a\Delta - \pi_2^a + 2\pi_2^a\Delta}$$

が成立し，結局

$$\hat{R}^\pi = 1/2R_1^\pi + 1/2R_2^\pi$$

が得られる．

$\hat{R}^\pi$ を最大化するためには，$\hat{R}^\pi = \hat{R}^\pi(\pi_1^a, \pi_2^a)$ の右辺を最大化すればよい．**図2**は，$\Delta = 0.01$ のときに，$\hat{R}^\pi(\pi_1^a, \pi_2^a)$ をプロットしたものである．π_1^a π_2^a 平面上の単位正方形が確率的政策の空間に対応し，単位正方形の4頂点が，決定的政策に対応する．図から，最適政策が決定的政策ではなく，確率的政策であることが読み取れる．実際，4つの決定的政策の平均報酬の期待値は0.5であるのに対して，最適な確率的政策の平均報酬の期待値は0.7である．

2.3 不確実なモデル問題の難しさ

不確実なモデル問題はその見掛けよりもずっと難しい問題であることが以下のようにして分かる．

まず，各モデル M_i に対して，政策 π のもとでの状態 s の（バイアス）値関数 $V_i^\pi(s)$ を以下の式を用いて定義する[6]．

$$V_i^\pi(s) + R_i^\pi = \sum_{a \in A} \pi(a|s) \times \sum_{s' \in S} P_i(s, a, s') \left(R_i(s, a, s') + V_i^\pi(s') \right) \quad (5)$$

また，（バイアス）Q値関数 $Q_i^\pi(s, a)$ を以下の式を用いて定義する．

$$Q_i^\pi(s, a) + R_i^\pi = \sum_{s' \in S} P_i(s, a, s')(R_i(s, a, s') + V_i^\pi(s')) \quad (6)$$

なお，式(5)と(6)から，以下が得られる．

$$\sum_a \pi(a|s) Q_i^\pi(s, a) = V_i^\pi(s) \quad (7)$$

このとき，以下の補題が成立する．

補題1 政策を π から π'（π，π' は任意）へ変更したとき，各モデル $\{M_i \mid 1 \le i \le n\}$ の平均報酬の変化量 $R_i^{\pi'} - R_i^\pi$ について，次式が成立する．

$$R_i^{\pi'} - R_i^\pi = \sum_{s \in S} \sum_{a \in A} \pi'(a|s) \left\{ P_i^{\pi'}(s)(Q_i^\pi(s, a) - V_i^\pi(s)) \right\}$$

証明を付録A.1に示した．この式から，完全観測マルコフ決定過程（すなわち，候補モデルが1つしかない場合）の政策改良手順[6]を導出できることに注目されたい．任意の s について，$a^* = \arg\max_a (Q_i^\pi(s, a) - V_i^\pi(s))$ とすれば，式(7)より $Q_i^\pi(s, a^*) - V_i^\pi(s) \ge 0$ であることが分かる．またエルゴード性の仮定より，任意の政策 π' について $P_i^{\pi'}(s) > 0$ である．したがって（決定的）政策 π' を各 s において $\pi'(a^*|s) = 1.0$ と設定すれば，$Q_i^\pi(s, a^*) - V_i^\pi(s) > 0$ なる状態とアクションの対 (s, a^*) が1つでもある限り，補題1の右辺の値は正となり，必ず平均報酬は増加する．

同様の式を平均報酬の期待値 $\hat{R}^\pi$ について導出できないか検討する．2章で定義した，各モデル M_i の政策 π のもとでの状態 s の状態占有確率 $P_i^\pi(s)$ の期待値を $\hat{P}^\pi(s)$ とすれば，

$$\hat{P}^\pi(s) = \sum_{i=1}^n q_i P_i^\pi(s) \quad (8)$$

$\hat{P}^\pi(s)$ はどのモデルにいるかを問わずに，政策 π のもとでセンサー入力 s が得られる確率と考えられる．また，政策が π，エージェントのセンサー入力が s のときに，真のモデルが M_i である確率を $\varphi_s^\pi(M_i)$ とすれば，ベイズの定理より，

$$\varphi_s^\pi(M_i) = \frac{Pr_\pi\{s \mid M_i\} Pr_\pi\{M_i\}}{\sum_{j=1}^n Pr_\pi\{s \mid M_j\} Pr_\pi\{M_j\}} = \frac{q_i P_i^\pi(s)}{\sum_{j=1}^n q_j P_j^\pi(s)} = \frac{q_i P_i^\pi(s)}{\hat{P}^\pi(s)} \quad (9)$$

すると，センサー入力が s であるときの値関数の期待値 $\hat{V}^\pi(s)$，Q 値関数の期待値 $\hat{Q}^\pi(s,a)$ は以下の式で計算できる．

$$\hat{V}^\pi(s) = \sum_{i=1}^{n} \varphi_s^\pi(M_i) V_i^\pi(s) \tag{10}$$

$$\hat{Q}^\pi(s,a) = \sum_{i=1}^{n} \varphi_s^\pi(M_i) Q_i^\pi(s,a) \tag{11}$$

なお，式 (7)，(10)，(11) から，以下が得られる．

$$\sum_a \pi(a|s)\hat{Q}^\pi(s,a) = \hat{V}^\pi(s) \tag{12}$$

しかし，残念ながら補題 1 と同様の式は平均報酬の期待値 $\hat{R}^\pi$ については成立しない．

補題 2 政策を π から π'（π，π' は任意）へ変更したとき，平均報酬の期待値について，次式が成立する．

$$\hat{R}^{\pi'} - \hat{R}^\pi = \sum_{s\in S}\sum_{a\in A} \pi'(a|s) \times \left\{ \hat{P}^{\pi'}(s) \sum_{i=1}^{n} \varphi_s^{\pi'}(M_i) \left(Q_i^\pi(s,a) - V_i^\pi(s)\right) \right\}$$

証明を付録 A.2 に示した．$\varphi_s^{\pi'}(M_i)$ の項が政策 π' に関するものであるために，$\sum_i$ 以降の項を $(\hat{Q}^\pi(s,a) - \hat{V}^\pi(s))$ と置き換えられないのである．

2.4 POMDP 問題との関連性

完全観測マルコフ決定過程では最適政策は決定的であり，モデルがエルゴード的であれば，値関数をすべての状態について同時に最大化することが知られている．しかし，不確実なモデル問題の最適政策は，たとえ各候補モデルが完全観測マルコフ決定過程であっても，このような望ましい性質を持たない．すでに上記の例題において，平均報酬の期待値を最大化する最適政策が決定的ではなく，確率的であることを見た．さらに，最適政策は値関数の期待値 $\hat{V}^\pi(s)$ をすべての s について，必ずしも同時に最大化しない．

不確実なモデル問題の持つこのような困難さは POMDP において典型的に現れる[14]ものであるが，この対応関係を次のようにして説明することができる．たとえ個々の候補モデルに対して状態が完全観測可能であっても，エージェントはどの候補モデルが真のモデルか分からないため，状態があたかも部分観測であるようにエージェントには見えるのである．このように考えれば，不確実なモデル問題を POMDP において研究されている知覚エイリアシング問題の一例と見なすことが可能になる．

POMDP の有限計画期間に対する最適政策を求める問題は *PSPACE* 完全であることが知られている[10]．無限計画期間に対する最適ポリシーを求める問題は，離散的問題ではなくなり，有限のプログラムで有限時間内に解くことは不可能である．Jaakkola ら[7]は POMDP における知覚エイリアシング問題を解くために，モンテカルロ法を用いた政策評価および政策改良のステップから構成されるモデルなし強化学習アルゴリズムを開発した．彼らは，POMDP における政策計算の困難さを軽減するために，政策を記憶なし政策に限定した．

不確実なモデル問題を解くために，筆者らは Jaakkola らの手法の変形版を開発した．その詳細については次章で説明する．不確実なモデル問題では，与えられた候補モデルとその不確実さから容易に政策を評価することができるので，モデルなし強化学習のように，計算量的に高価なモンテカルロ法を用いて政策評価を行う必要はない．

3. 不確実なモデル問題の解法

この章では 2 章で定義された不確実なモデル問題の解法について説明する．

3.1 平均報酬の期待値の勾配ベクトル

2.3 節で述べた不確実なモデル問題の困難さは，不確実性がない場合は線形計画問題であるのに対して，不確実性がある場合は非線型計画問題であるからともいえる．残念ながら補題 1 と同様の式は平均報酬の期待値 $\hat{R}^\pi$ については成立しないことが分かった．そこで非線型計画問題の常套手段である局所線形化を補題 2 の式について検討してみよう．

まず，次の補題 3 が成立することに着目する．

補題 3 $\forall\epsilon, \exists L > 0, \text{s.t. } \forall i\ (1 \le i \le n), \forall s \in S$

$$\max_{s\in S, a\in A} |\pi'(a|s) - \pi(a|s)| < \epsilon \Rightarrow \begin{cases} |\hat{P}^{\pi'}(s) - \hat{P}^\pi(s)| < L\epsilon \\ |\varphi_s^{\pi'}(M_i) - \varphi_s^\pi(M_i)| < L\epsilon \end{cases}$$

証明を付録 A.3 に示した．補題 3 は $\{\hat{P}^\pi(s), \varphi_s^\pi(M_i) \mid s \in S, 1 \le i \le n\}$ のそれぞれを変数を $\{\pi(a|s) \mid a \in A, s \in S\}$ とする多変数関数と見なしたとき，これらの関数がいわゆるリプシツ連続であることを示している．

さて，補題 2，補題 3 より，次の定理が成立することを証明できる．

定理 1 $\Delta\pi(a|s) = \pi'(a|s) - \pi(a|s)$ とおく．もし $\max_{s\in S, a\in A} |\Delta\pi(a|s)| < \epsilon$ であれば，

$$\hat{R}^{\pi'} - \hat{R}^\pi = \sum_{s\in S}\sum_{a\in A} \Delta\pi(a|s) \times \left\{ \hat{P}^\pi(s) \left(\hat{Q}^\pi(s,a) - \hat{V}^\pi(s) \right) \right\} + O(\epsilon^2)$$

付録 A.4 に証明を示した．この定理は平均報酬の期待値 $\hat{R}^\pi$ の政策 $\{\pi(a|s)|s \in S, a \in A\}$ に関する偏微分係数は $\hat{P}^\pi(s)(\hat{Q}^\pi(s,a) - \hat{V}^\pi(s))$ であることを示している．これらの偏微分係数を適当にならべて勾配ベクトルを得ることができる．

3.2 PIMCM アルゴリズム

2 章で定義された不確実なモデル問題を解決するために考案した PIMCM（Policy Iteration for Multiple Candidate Models）アルゴリズムについて説明する．PIMCM アルゴリズムは定理 1 に基づいて，確率的政策の空間で山登り探索を行う．

Step 1 ϵ を十分小さい値に設定する．初期政策 π を任意に設定する．

Step 2 それぞれのモデル候補 $\{M_i \mid 1 \leq i \leq n\}$ について，政策 π を評価し，$P_i^\pi(s)$，R_i^π，および $V_i^\pi(s)$，$Q_i^\pi(s,a)$ を計算する．

Step 3 前ステップの計算結果とモデル候補 M_i の確率 q_i から $\hat{P}^\pi(s)$，$\varphi_s^\pi(M_i)$，$\hat{V}^\pi(s)$，$\hat{Q}^\pi(s,a)$ を計算する．

Step 4 以下が成立するか否か調べる．

$$\exists s \in S \quad \text{s.t.} \max_a \hat{Q}^\pi(s,a) > \hat{V}^\pi(s)$$

もし成立しないならば，反復を終了し π を返す．

Step 5 それぞれの s について，$a^* = \arg\max_a \hat{Q}^\pi(s,a)$ として，確率的政策 $\pi^1(a|s)$ を次のように定める．

$$\pi^1(a|s) = 1.0 \ (a = a^*)$$
$$\pi^1(a|s) = 0.0 \ (a \neq a^*)$$

次に，確率的政策 π^ϵ を以下のように定める．

$$\pi^\epsilon(a|s) = (1-\epsilon)\pi(a|s) + \epsilon\pi^1(a|s)$$

Step 6 $\pi = \pi^\epsilon$ として，Step 2 へもどる．

なお，このアルゴリズムの計算量は $|A|$ を定数と考えれば，Step 2 で $O(n|S|^3)$，Step 3 で $O(n|S|)$ である．

3.3 PIMCM アルゴリズムの収束性

PIMCM アルゴリズムが平均報酬の期待値の局所最大を見つけることができることは，次の定理により証明できる．

定理 2 もし ϵ が十分小さく設定されていれば，PIMCM アルゴリズムの政策変更（Step 5）は，実際は政策改良である．また，PIMCM アルゴリズムが終了する（Step 4）ときには，その政策における勾配ベクトルは零ベクトルである．

（証明）

PIMCM アルゴリズムの Step 5 が政策改良であることは以下のように示せる．まず，次式を定理 1 に代入する．

$$\begin{aligned}\Delta\pi(a|s) &= \pi^\epsilon(a|s) - \pi(a|s) \\ &= \epsilon(\pi^1(a|s) - \pi(a|s))\end{aligned}$$

すると次式が得られる．

$$\hat{R}^{\pi^\epsilon} - \hat{R}^\pi = \epsilon \sum_s \sum_a (\pi^1(a|s) - \pi(a|s)) \times \left\{\hat{P}^\pi(s)(\hat{Q}^\pi(s,a) - \hat{V}^\pi(s))\right\} + O(\epsilon^2)$$

式 (12) により，

$$\sum_s \sum_a \pi(a|s)\{\hat{P}^\pi(s)(\hat{Q}^\pi(s,a) - \hat{V}^\pi(s))\} = 0$$

なので，結局

$$\hat{R}^{\pi^\epsilon} - \hat{R}^\pi = \epsilon \sum_s \sum_a \pi^1(a|s) \times \left\{\hat{P}^\pi(s)(\hat{Q}^\pi(s,a) - \hat{V}^\pi(s))\right\} + O(\epsilon^2)$$

が得られる．ϵ が十分小さければ，上式の値は正である．

PIMCM アルゴリズムが終了したときに勾配ベクトルが 0 ベクトルであること，すなわちすべての $s \in S$，すべての $a \in A$ について $\hat{P}^\pi(s)(\hat{Q}^\pi(s,a) - \hat{V}^\pi(s)) = 0$ であることは次のように示せる．勾配ベクトルに正の要素があるときには明らかにアルゴリズムは終了しない．式 (12) から，$\hat{V}^\pi(s)$ は $\{\hat{Q}^\pi(s,a)|a \in A\}$ の重みつき平均値であることが分かるので，勾配ベクトルに負の要素があるときには必ず正の要素も存在する．したがって，アルゴリズムが終了するのは，勾配ベクトルのすべての要素が 0 の場合のみである． □

勾配が零ベクトルになるのを停留条件と呼び，そのような点を停留点と呼ぶ．得られた政策の局所最大性を厳密に示すには，停留条件だけでなく，2 次偏微分の項を調べるなどして，停留点の近傍において平均報酬の期待値が上に凸であることを示す必要がある．しかし，PIMCM アルゴリズムではその検討を行っていない．一般に最急降下法などの非線型計画法アルゴリズムにおいても，局所最適解でない停留点に収束する場合はまれであり，仮にあっても，変数に小さな摂動を与えて再出発すればほぼ確実にその点から離れていくとして，停留条件のみを調べているようである[8]．

4. 応用問題

応用問題に適用することにより，PIMCM アルゴリズムの有用性を示す．

4.1 繰返し囚人のジレンマ問題

繰返し囚人のジレンマ問題とは**図 3** に示した利得表の双行列ゲームを繰り返し実行するものである．この問題では，アクション集合 $A = \{c, d\}$ である．ここに c は協力を，d は裏切りを表す．ここで考える繰返し囚人のジレンマ問題では，エージェントは，自分自身，お

プレイヤー1＼プレイヤー2	協力 (c)	裏切り (d)
協力 (c)	(3,3)	(0,5)
裏切り (d)	(5,0)	(1,1)

図3 囚人のジレンマ問題の利得表．各欄は，対応するアクション選択における（プレイヤー1，プレイヤー2）の利得を示す．

Fig. 3 Payoffs for the Prisoner's Dilemma. Each item shows payoffs for (player1, player2) under the corresponding action selections.

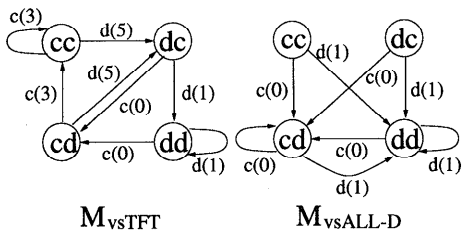


図4 繰返し囚人のジレンマ問題．$\hat{M} = \{M_{vsTFT}, M_{vsALL-D}\}$，$\hat{q} = (1/2, 1/2)$．2つのMDPモデルをエルゴード的にするために，1%のアクション・エラーを入れた．

Fig. 4 Iterated Prisoner's Dilemma. $\hat{M} = \{M_{vsTFT}, M_{vsALL-D}\}$, $\hat{q} = (1/2, 1/2)$. We put 1% action error to make the MDPs ergodic.

よび対戦相手の前回のアクションを記憶していると仮定する．したがって，状態空間は $S = \{cc, cd, dc, dd\}$ となる．たとえば，状態 cd は，自分の前回のアクションが c，対戦相手の前回のアクションが d である状態を表すものとする．また，対戦相手は，TFT（しっぺ返し）プレイヤーと，ALL-D（全面裏切り）プレイヤーの2人である．TFTプレイヤーは，始めはアクション c をとり，それ以降は相手の前回とったアクションをとるものであり，Axelrodの行った2度のトーナメントで優勝したことでよく知られる[1]．エージェントは，TFTプレイヤー，ALL-Dプレイヤーのうち等確率で選ばれたものと対戦しなければならない．ただし，エージェントは対戦相手がTFTプレイヤーかALL-Dプレイヤーのどちらか識別できず，したがって，2種類のプレイヤーに対して同じ政策を用いなければならないと仮定する．対戦相手の政策が既知の場合，繰返し囚人のジレンマ問題は，1つの既知のMDPを解くことに帰着される．しかし，今，対戦相手が2者のうちどちらかが不確実であるため，対戦相手がTFTである場合とALL-Dである場合の2つのMDPモデルからなる集合 $\hat{M} = \{M_{vsTFT}, M_{vsALL-D}\}$ 上の不確実なモデル問題となる．以上の問題設定を**図4**に示す．

この不確実なモデル問題 $(S, A, \hat{M}, \hat{q})$ にPIMCMアルゴリズムを適用して見つかった政策は，$\pi^*(c|cc) = 1.0$，$\pi^*(c|cd) = 0.3$，$\pi^*(c|dc) = 1.0$，$\pi^*(c|dd) = 0.1$ である．この政策の平均報酬の期待値 $(\hat{R}^{\pi^*})$ は，1.83である．一方，16種類の決定的政策の平均報酬の期待値は，最高で1.64，最低で0.98，平均で1.38である．

4.2 モデルベースの強化学習

モデルベースの強化学習法では，環境のモデルを学習により構築し，構築されたモデルを解くことにより優れた政策を獲得することを目指す．そのような手法では，獲得されたモデルに不確実性がともなうが，これまで不確実なモデル問題を扱うことができなかったため最も確からしいモデルを解くことで政策を得ていた．しかし，そのようにして得られた政策は，必ずしも真の環境での優れた政策ではなく，実際には著しく低い報酬しか達成できないかもしれない．そこで，単に最も確からしいモデルを解く代わりに，複数のモデル候補集合上の不確かなモデル問題としてとらえ，PIMCMアルゴリズムを用いて政策を求めれば，そのような危険性を低く抑えることができる．

ここでは，モデルベースの強化学習法として，BLHT (Bayesian learning of history trees)[15),16)]を用いる．BLHTでは，環境はヒストリーツリーと呼ばれる，可変長履歴を持つ確率モデルでモデル化され，候補モデルの集合は，通常は履歴の最大長さ d により，指定される．BLHTの長所の1つは，モデル候補の数が膨大になっても☆，効率良くモデル学習できる点にある．ヒストリーツリーからなる候補モデルの集合を与えられたとすれば，BLHTのモデル学習は以下のように進められる．

(1) 環境を探検し，個々のモデル候補の事後確率を計算する．
(2) 探検を終了する．MAPモデル（事後確率最大のモデル）を取り出し，MAPモデルに対する最適政策を計算する．

したがって，この応用例で扱うMDPモデルは，長さ d の履歴が状態（センサー入力）であり，得られる政策は，各履歴に対するアクション集合上の確率分布である．

実験に用いた環境を**図5**に示す．この環境には5つの状態があり，各状態でエージェントはそれぞれ図に示されたセンサー入力を得る．また，エージェントのとれるアクションはa，bのいずれかである．図から分かるように，S_1，S_2 というセンサー入力を得る状態はそれぞれ2つある．つまり，知覚エイリアシングがあり，POMDP問題となっている．

☆ モデル候補の数は履歴の最大長さ d の指数関数である．

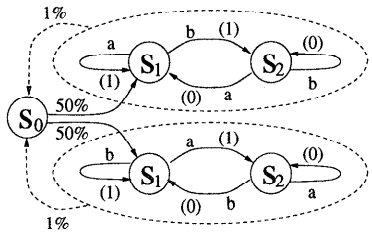


図5 BLHTが学習する環境．2つの破線で囲まれた4つの状態からは，すべてのアクションについて，1%の確率で，状態 S_0 へ推移する．

Fig. 5 Environment for BLHT. There is 1% random transition to S_0 from each of the four states in two dotted circles, no matter which action is taken.

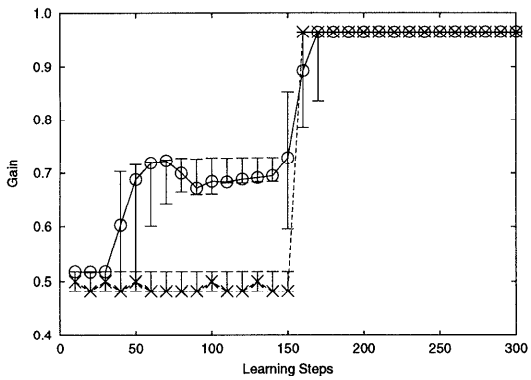


図6 BLHTにより得られた政策の平均報酬（破線）とPIMCMにより得られた政策の平均報酬（実線）．10回の試行の中央値を示した．ただし，エラーバーは，25–75パーセンタイルの区間を示す．

Fig. 6 The gain of the policy by BLHT (in a dashed line) and the gain of the policy by PIMCM (in a solid line). The medians of ten trials are plotted. Error-bars shows 25–75 percentile intervals.

BLHTがMAPモデルに対してダイナミック・プログラミングを用いて求めた決定的政策の平均報酬と，そのMAPを含む事後確率の最も大きい3つのモデル候補の集合上の不確実なモデル問題へPIMCMアルゴリズムを適用して求めた☆確率的政策の平均報酬を**図6**に示す．平均報酬はBLHTの学習ステップ数の関数としてプロットした．PIMCMアルゴリズムと組み合わせた場合，学習の比較的初期段階で得られる政策が大幅に改善されていることが分かる．

☆ 履歴の最大長さ $d = 7$ と設定したので，候補モデルの総数は約 4×10^{75} になった．候補モデルの総数が非常に大きいため，MAPモデルでさえも，その事後確率は非常に小さい．したがって，3つのモデル候補については，同じ事後確率であるとして，PIMCMアルゴリズムを適用した．

5. 関連研究

不確実なモデル問題に関連した研究については，すでに1章で説明した．不確実なモデル問題に一見よく似た問題にMDPIPsがある．

MDPIPs（MDPs with Imprecise Parameters）は遷移確率関数，報酬関数を表すパラメータが正確には知られていないMDPである．MDPIPに対する最良の政策を計算する問題は，本論文で考察した不確実なモデル問題とは異なる問題である．MDPIP問題においては候補モデルは1つしかない．MDPIP問題のベイズ定式化においては，各時間ステップごとに，パラメータはその確率分布に従ってランダムに選ばれ，平均報酬のパラメータの分布に対する期待値を最大化する政策がベイズ最適政策として計算される．したがって，もし試行結果からパラメータの確率分布を更新するような学習過程を取り入れなければ，MDPIP問題のベイズ定式化は，パラメータの平均値を計算することにより，パラメータが正確に与えられた通常のMDP問題に帰結することができる．なお，MDPIP問題の非ベイズ定式化は，Satia と Lave[12]，Givan ら[5]により，マックス・ミン戦略，およびマックス・マックス戦略を用いて研究されている．

6. まとめ

アクションの効果が決定的ではなく，確率的であるような問題領域での計画立案について，マルコフ決定過程理論などに基づく決定理論的計画立案が提案されている．本論文では，これまでの決定理論的計画立案研究では考慮されていなかったモデルの不確実性を考慮して政策を計算するアルゴリズムを説明した．

候補モデルの有限集合とその集合上の確率分布（確信度）が与えられると仮定すれば，最適政策は平均報酬の期待値を最大化する政策として定義できる．筆者らは，平均報酬の期待値を局所最大化するように，確率的政策を改良するための，政策反復アルゴリズム（PIMCMアルゴリズム）を開発した．

不確実なモデル問題は，強化学習の研究でよく知られているモデルの「探検対利用問題」[17]（the exploitation vs exploration problem）とも関連している．ベイズ学習アルゴリズムとPIMCMアルゴリズムを組み合わせて，より優れた探検戦略を求めることが，今後の研究課題である．

謝辞 本研究で開発したアルゴリズムの実装，実験の実施は同じ研究室の前谷真二君によるものである．同君の協力に深く感謝する．

参考文献

1) Axelrod, R.: *The Evolution of Cooperation*, Basic Books (1984).
2) Ben-Porath, E.: The Complexity of Computing a Best Response Automaton in Repeated Games with Mixed Strategies, *Games and Economic Behavior*, Vol.2, pp.1–12 (1990).
3) Boutilier, C., Dearden, R. and Goldszmidt, M.: Exploiting Structure in Policy Construction, *Proc. 14th International Joint Conference on Artificial Intelligence*, pp.1104–1111 (1995).
4) Dean, T., Kaelbling, L.P., Kirman, J. and Nicholson, A.: Planning with deadlines in stochastic domains, *Proc. 11th National Conference on Artificial Intelligence*, pp.574–579 (1993).
5) Givan, R., Leach, S. and Dean, T.: Bounded Paramctcr Markov Decison Processes, *Proc. 4th European Conference on Planning (ECP '97)* (1997).
6) Howard, R.A.: *Dynamic Programming and Markov Processes*, MIT Press (1960). 関根智明, 羽鳥裕久, 森 俊夫 (訳): ダイナミックプログラミングとマルコフ過程, 培風館 (1971).
7) Jaakkola, T., Singh, S.P. and Jordan, M.I.: Reinforcement Learning Algorithm for Partially Observable Markov Decision Problems, *Advances of Neural Information Processing Systems 7*, pp.345–352 (1994).
8) 今野 浩, 山下 浩: 非線形計画法, 日科技連 (1978).
9) 森村英典, 高橋幸雄: マルコフ解析, 日科技連 (1979).
10) Papadimitriou, C. and Tsitsiklis, J.: The Complexity of Markov Decision Problems, *Mathematics of Operations Research*, Vol.12, No.3, pp.441–450 (1987).
11) Puterman, M.L.: *Markov Decision Processes*, Wiley (1994).
12) Satia, J.K. and Lave, R.E.: Markovian Decision Processes with Uncertain Transition Probabilities, *Operations Research*, Vol.21, pp.728–740 (1973).
13) Schneider, J.G.: Exploiting Model Uncertainty Estimates for Safe Dynamic Control Learning, *Advances of Neural Information Processing Systems 9*, pp.1047–1053 (1996).
14) Singh, S., Jaakkola, T. and Jordan, M.: Learning Without State-Estimation in Partially Observable Markovian Decision Processes, *Proc. 11th International Conference on Machine Learning*, pp.284–292 (1994).
15) 末松伸朗, 林 朗, 李 仕剛: 部分観測環境での強化学習へのモデルベースアプローチ: 可変長記憶モデルのベイズ学習, 人工知能学会誌, Vol.13, No.3, pp.404–414 (1998).
16) Suematsu, N., Hayashi, A. and Li, S.: A Bayesian approach to model learning in non-Markovian environments, *Proc. 14th International Conference on Machine Learning*, pp.349–357 (1997).
17) Sutton, R.S. and Barto, A.G.: *Reinforcement Learning*, MIT Press (1997).

付 録

A.1 補題1の証明

式 (6) より,

$$R^{\pi'} - R^{\pi} = \sum_{s'} P_i(s,a,s')\Big(V_i^{\pi'}(s') - V_i^{\pi}(s')\Big) - \Big(Q_i^{\pi'}(s,a) - Q_i^{\pi}(s,a)\Big)$$

上式に $\pi'(a|s)$ をかけて $a \in A$ について合計する. $\sum_a \pi'(a|s) = 1.0$ なので,

$$R^{\pi'} - R^{\pi} = \sum_a \pi'(a|s)\Big\{\sum_{s'} P_i(s,a,s') \times \Big(V_i^{\pi'}(s') - V_i^{\pi}(s')\Big) - \Big(Q_i^{\pi'}(s,a) - Q_i^{\pi}(s,a)\Big)\Big\}$$

が得られる. 上式に $P_i^{\pi'}(s)$ をかけて $s \in S$ について合計する. $\sum_s P_i^{\pi'}(s) = 1.0$ なので,

$$R^{\pi'} - R^{\pi} = \sum_s P_i^{\pi'}(s) \sum_a \pi'(a|s) \times \Big\{\sum_{s'} P_i(s,a,s')(V_i^{\pi'}(s') - V_i^{\pi}(s')) - (Q_i^{\pi'}(s,a) - Q_i^{\pi}(s,a))\Big\}$$

この式の右辺を展開する. $\sum_s P_i^{\pi'}(s) \sum_a \pi'(a|s) \sum_{s'} P_i(s,a,s') = \sum_{s'} P_i^{\pi'}(s')$ を用いると,

$$1\text{項} = \sum_s P_i^{\pi'}(s) \sum_a \pi'(a|s) \times \sum_{s'} P_i(s,a,s') V_i^{\pi'}(s') = \sum_s P^{\pi'}(s) V_i^{\pi'}(s)$$

$$2\text{項} = -\sum_s P_i^{\pi'}(s) \sum_a \pi'(a|s) \times \sum_{s'} P_i(s,a,s') V_i^{\pi}(s')$$

$$= -\sum_s P^{\pi'}(s)V_i^{\pi}(s)$$

$$3\text{項} = -\sum_s P_i^{\pi'}(s)\sum_a \pi'(a|s)Q_i^{\pi'}(s,a)$$

$$4\text{項} = \sum_s P_i^{\pi'}(s)\sum_a \pi'(a|s)Q_i^{\pi}(s,a)$$

式(7)が成立することに着目すると，上の式の第1項と第3項はキャンセルしあい，第2項と第4項が残り，

$$R_i^{\pi'} - R_i^{\pi} = \sum_s\sum_a \pi'(a|s)P_i^{\pi'}(s)(Q_i^{\pi}(s,a) - V_i^{\pi}(s))$$

が得られる． □

A.2 補題2の証明

補題1の等式の両辺に q_i を書けて $1 \le i \le n$ について合計する．左辺を平均報酬の期待値の定義式(4)を用いて書き直し，右辺の総和の順序を変えると，

$$\hat{R}^{\pi'} - \hat{R}^{\pi} = \sum_s\sum_a \pi'(a|s) \times \left\{\sum_i q_i P_i^{\pi'}(s)(Q_i^{\pi}(s,a) - V_i^{\pi}(s))\right\}$$

が得られる．式(9)により，右辺の $q_iP_i^{\pi'}(s)$ を $\hat{P}^{\pi'}(s)\varphi_s^{\pi'}(M_i)$ と書き直せば，補題2が得られる． □

A.3 補題3の証明

確率的政策のなす集合 Π を

$$\Pi = \{\pi(a_j|s) \mid 1 \le j \le |A|-1, s \in S, \pi(a_j|s) \ge 0, \sum_{j=1}^{|A|-1}\pi(a_j|s) \le 1.0\}$$

とする．Π を，$S(|A|-1)$ 次元ユークリッド空間の部分集合と考えると，Π はコンパクト集合である．

また，すべての i について，状態占有確率 $\{P_i^{\pi}(s) \mid s \in S\}$ は式(1)と式(2)を連立方程式として解いた解であり，クラメールの公式から Π 上の C^1 級関数であることが分かる．すると定義より，$\{\hat{P}^{\pi}(s) \mid s \in S\}$，$\{\varphi_s^{\pi}(M_i) \mid s \in S, 1 \le i \le n\}$ のそれぞれも Π 上の C^1 級関数であることが分かる．

解析学の定理により，コンパクト集合上の C^1 級関数はリプシツ連続である．したがって，$\{\hat{P}^{\pi}(s) \mid s \in S\}$，$\{\varphi_s^{\pi}(M_i) \mid s \in S, 1 \le i \le n\}$ のそれぞれに対して補題を満たすような正の定数 L が存在する．それらの最大値をあらためて L とおいてやればよい． □

A.4 定理1の証明

以下の式を補題2に代入する．

$$\pi'(a|s) = \pi(a|s) + \Delta\pi(a|s)$$
$$\hat{P}^{\pi'}(s) = \hat{P}^{\pi}(s) + \Delta\hat{P}^{\pi}(s)$$
$$\varphi_s^{\pi'}(M_i) = \varphi_s^{\pi}(M_i) + \Delta\varphi_s^{\pi}(M_i)$$

補題3により，$\Delta\hat{P}^{\pi}(s)$，$\Delta\pi(a|s)$，$\Delta\varphi_s^{\pi}(M_i)$ が $O(\epsilon)$ であるので，

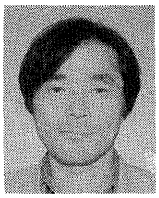
$$\begin{aligned}\hat{R}^{\pi'} - \hat{R}^{\pi} &= \sum_s\sum_a \pi(a|s)\hat{P}^{\pi}(s) \times \sum_i \varphi_s^{\pi}(M_i)\,(Q_i^{\pi}(s,a) - V_i^{\pi}(s)) \\ &+ \sum_s\sum_a \Delta\pi(a|s)\hat{P}^{\pi}(s) \times \sum_i \varphi_s^{\pi}(M_i)\,(Q_i^{\pi}(s,a) - V_i^{\pi}(s)) \\ &+ \sum_s\sum_a \pi(a|s)\Delta\hat{P}^{\pi}(s) \times \sum_i \varphi_s^{\pi}(M_i)(Q_i^{\pi}(s,a) - V_i^{\pi}(s)) \\ &+ \sum_s\sum_a \pi(a|s)\hat{P}^{\pi}(s) \times \sum_i \Delta\varphi_s^{\pi}(M_i)(Q_i^{\pi}(s,a) - V_i^{\pi}(s)) \\ &+ O(\epsilon^2)\end{aligned}$$

式(7)により，第1項，第3項，第4項は0となる．第2項に式(10)と式(11)を代入すれば，定理1が得られる． □

(平成10年8月26日受付)
(平成11年4月 1 日採録)

末松 伸朗（正会員）

1988年九州大学理学部物理学科卒業．1990年同大学院修士課程修了．同年，(株) 富士通研究所入社．1994年広島市立大学情報科学部助手，現在に至る．機械学習，知能ロボットの研究に従事．人工知能学会，日本認知科学会各会員．



林　　朗（正会員）

1974年京都大学理学部数学科卒業．同年，日本アイビーエム（株）入社．構造解析ソフトウェアの開発に従事．1988年ブラウン大学計算機科学科修士課程修了．1991年テキサス大学オースチン校計算機科学科博士課程修了（Ph.D.）．九州工業大学情報工学部客員助教授を経て，現在は広島市立大学情報科学部情報機械システム工学科教授．研究対象はロボット学習，空間推論，ゲーム理論．人工知能学会，日本ロボット学会，日本機械学会，AAAI，ACM，IEEE各会員．