解　説

# 部分観測マルコフ決定過程下での強化学習

## Reinforcement Learning for Partially Observable Markov Decision Processes

木村　　元[*1]　Leslie Pack Kaelbling[*2]
Hajime Kimura

＊1　東京工業大学　大学院総合理工学研究科
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology.
＊2　Dept. of Computer Science, Brown University, USA.

19YY 年 MM 月 DD 日　受理

**Keywords:** reinforcement learning, POMDP, Markov decision process, partial observability, belief state

## 1. はじめに

部分観測マルコフ決定過程 (Partially Observable Markov Decision Process:POMDP) における強化学習は，環境に不確実性があるという仮定を出発点としている点において有望な適応的な学習制御の枠組である．多くの実世界の制御問題では，不確実性の扱いはエンジニアにとって頭痛の種である．行動や評価にノイズが存在したり，不十分なセンサのため状態観測に不確実性や不完全性の存在するシステムのモデルとして，POMDP は特に適している．POMDP の理論は不確実性のもとでの最適なふるまいを求めるための基礎となる．[Simmons et al. 95] や [Cassandra et al. 96] は，自動車のナビゲーションにおける位置認識の不確実性を扱うために POMDP のモデルを利用し，共によい結果を得ている．

強化学習もまた，試行錯誤の経験を通して，遅れのある報酬を手がかりにしてシステムの挙動を改善する適応的学習制御の枠組として興味深い．強化学習を用いることにより，システムは設計者の知識の欠落を補ったり，環境の変化に自ら追従できるため，人工知能の基本的研究分野の一つとなりつつある．多くの強化学習とその理論的な解析では，環境をマルコフ決定過程 (Markov decision process: MDP) としてモデルしており，状態の観測は完全であることを仮定していた．POMDP における強化学習の厳密解法は、残念ながら極端に小さいかあるいは複雑さの小さい問題を除き，計算量的に実行不可能と考えられている．よって，POMDP 環境下での強化学習問題は非常に意欲的な課題であると考えられる．

本論文は，POMDP における強化学習の研究について概説する．まず，POMDP モデルについて解説し，問題を解く手がかりとなる重要な特徴について述べる．次に，POMDP におけるいくつかの代表的な強化学習法を紹介し，POMDP のどのような特徴に依存しているのかについて考察する．最後に，紹介した手法の利点や欠点について総合的に考察し，今後の研究の動向について考える．

## 2. 問題設定

### 2・1　マルコフ決定過程 ( MDP )

離散時間の MDP は，状態集合 $S$ と行動集合 $A$ によって構成される．各時間ステップ $t$ において，まずエージェントは現在の状態 $s_t$ を観測し，行動 $a_t$ を実行し，環境から直接報酬 $r_t$ を受け取る．これでそのステップでのインタラクションは終了し，次の時間ステップへ進む．環境は確率 $P^a_{ss'}$ に従って状態遷移を行う．ただし $P^a_{ss'}$ は，状態 $s$ において行動 $a$ を実行したとき $s'$ へ遷移する確率を表す．報酬も確率変数だが，直接報酬の期待値 $R^a(s)$ は現在の状態と行動のみに依存する．マルコフ性の仮定により，状態遷移確率と報酬は状態 $s_t$ よりも以前の状態に対して独立である．

ある決定的な定常政策 $\pi$: $S \to A$ は，それぞれの状

態における行動を割り当てるような写像を表す．エージェントの学習の目的は，パフォーマンスを最大化する最適な政策を求めることである．最適性の評価規範として，無限期間の割引報酬の合計を用いることが多い．ある政策 $\pi$ において，Value は以下のように与えられる．

$$V_\infty(s) = E\left\{\sum_{t=0}^{\infty} \gamma^t r_t \,|s_0 = s\right\} ,$$

ただし $\gamma$, $0 \leq \gamma \leq 1$ は割引率と呼ばれ，将来の報酬に対してどの程度割り引いて評価するかを決めるパラメータである．すべての定常政策 $\pi$ は以下の式を満たす．

$$V_\infty^\pi(s) = R^{\pi(s)}(s) + \gamma \sum_{s' \in S} P_{ss'}^{\pi(s)} V_\infty^\pi(s') ,$$

ただし $\pi(s)$ は政策 $\pi$ のもとで状態 $s$ において選択する行動を表す．すべての状態 $s \in S$ において Value function を最大化する政策を最適政策という．すべての有限マルコフ決定過程において，最適な決定的定常政策が少なくとも 1 つ存在し，以下のような最適な Value function が 1 つだけ存在する [Howard 60]．for all $s \in S$，

$$V_\infty^*(s) = \max_a \left( R^a(s) + \gamma \sum_{s' \in S} P_{ss'}^a V_\infty^*(s') \right) .$$

最適政策は最適な Value function より容易に得ることができる．

状態 $s$ において行動 $a$ を実行した後，ずっと最適政策をとり続ける場合の無限期間の割引報酬の期待値を $Q_\infty^*(s,a)$ と表す．これも同様に以下のような式で表すことができる．

$$\begin{aligned} Q_\infty^*(s,a) &= R^a(s) + \gamma \sum_{s' \in S} P_{ss'}^a \max_{a' \in A} Q_\infty^*(s',a') \\ &= R^a(s) + \gamma \sum_{s' \in S} P_{ss'}^a V_\infty^*(s'). \end{aligned}$$

最適な Q 値が与えられると，最適政策はそれぞれの状態 $s$ において最大の $Q_\infty^*(s,a)$ を持つ行動に一致する．

### 2・2 MDP の解法と計算量

MDP のモデルすなわち状態遷移確率と直接報酬の関数が与えられると，ダイナミックプログラミング (DP) に基づく様々なアルゴリズムを用いて問題を解くことができる．Value iteration 法や policy iteration 法はよく知られている．MDP は多項式時間で解けることが示されている [Littman et al. 95b]，[Papadimitriou et al. 87]．

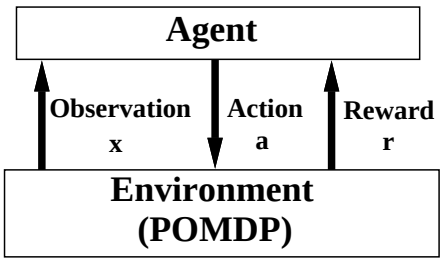



図 1　POMDP の環境とエージェントとの相互作用．

### 2・3 部分観測性の付加

POMDP は，エージェントの状態観測に不確実性を付加することにより MDP を拡張したものである．有限な観測の集合を $X$ と記す．各時間ステップ $t$ において，エージェントは状態 $s$ を直接観測できない代わりに，$P(x_t|s_t)$ の確率で $x_t \in X$ を観測する (図 1)．直接報酬より得られる情報も観測の中に明示的に含めることも可能である．エージェントが環境に関する完全なモデルと全ての過去の経験を保持しているとしても，部分観測性が存在するためにエージェントは現在の状態を完全に確認することは不可能である．しかしながら，エージェントが環境中のどの状態にいるのかを表す確率分布を求めることは可能である．この確率分布は信念 (belief) と呼ばれる．ある信念 $b$ はベクトルで表され，それぞれの要素はエージェントの現在の状態が $s$ である確率を表す $b(s)$ で構成される．

POMDP の例として，エージェントが 2 つのドアの前にいる状況を考える [Kaelbling et al.]．片方のドアの向こうの部屋にはトラがおり，もう片方のドアの向こうには大きな正の報酬がある．エージェントがトラのいる方のドアを開けてしまうと，トラに襲われて大きな負の報酬を受け取る「ドアを開ける」という行動の他に，トラがどちらにいるのかを知るために「トラの鳴き声を聞く」という行動をとることもできるが，その行動にはコスト ( 負の報酬 ) がかかり，その上それによって得られる情報も正確ではない．つまり，右側のドアから鳴き声が聞こえてきても，実際にはトラは左側にいる場合もあるという具合である．この例題の環境では，トラが右にいる状態を $s_r$，左にいる状態を $s_l$ と記し「左のドアを開ける」行動を $a_1$「トラの鳴き声を聞く」行動を $a_2$「右のドアを開ける」行動を $a_3$ とする．観測入力として「何もわからない」を $x_0$，

「右側から鳴き声が聞こえる」を $x_1$,「右側から鳴き声が聞こえる」を $x_2$，と記す．エージェントが行動 $a_2$ を実行した直後の観測は $x_1$ または $x_2$ のどちらかである．エージェントの最初の時間ステップでは $x_0$ を観測する．環境の状態遷移規則を以下に示す．エージェントがドアを開けて報酬あるいは罰を受け取った直後，環境はリセットされ，再びトラはどちらかのドアの向こうにランダムに配置され，この直後のエージェントの観測は $x_0$ となる．よって，行動 $a_1$ または $a_3$ を実行することによって環境は状態 $s_r$ または $s_l$ へ確率 0.5 で遷移する．トラのいる方のドアを開けた場合の報酬は $-100$ で，別のドアを開けると報酬は $+10$ である．「トラの鳴き声を聞く」行動 $a_2$ を実行すると，報酬は $-1$ となるが，環境の状態は変化しない．トラが右側にいる状態 $s_r$ において「トラの鳴き声を聞く」行動 $a_2$ を実行すると，確率 0.85 で「右側から鳴き声が聞こえる」$x_1$ を観測し，確率 0.15 で「左側から鳴き声が聞こえる」$x_2$ を観測する．トラが左側にいる場合は逆である．このようなトラを避ける問題を例にあげながら，POMDP の有用な特徴について以下に列挙する．

〔 1 〕 信念 (belief) は過去の履歴の縮約である

信念 (belief) は，現在の時点に至るまでのすべての観測と行動の履歴および初期状態分布を利用して行動決定する場合において，十分な統計量を持っている．つまり，現時点の信念状態が与えられれば，過去の履歴を利用してそれ以上の報酬を得るような行動を得ることはできない．POMDP の完全なモデルすなわち状態遷移確率 $P^a_{ss'}$ および観測の確率 $P(x|s)$ が与えられていれば，現時点の観測 $x_t$ と 1 ステップ前の信念を用いて，エージェントは常に信念を正しく計算して維持できる．トラの例題では，最初トラがどちらにいるか全く分からないので，信念 $b(s_r) = 0.5, b(s_l) = 0.5$ だが「トラの鳴き声を聞く」行動 $a_2$ を実行した結果「右側から鳴き声が聞こえる」$x_1$ を観測したならば，確率 0.85 でトラが右側にいると考えられるので信念 $b(s_r) = 0.85, b(s_l) = 0.15$ となる．

〔 2 〕 信念空間の **MDP(Belief MDP)**

POMDP は，信念 (belief) を状態空間とした MDP の問題 (belief MDP) へ帰着される．信念状態空間は，一般に不加算連続な空間である．

〔 3 〕 凸型の **Value Function**

信念状態空間における MDP(belief MDP) においても，通常の MDP と同様に value function と Q-function を定義できる．ここで，これらの関数のたて軸を value や Q 値とおき，軸の上方向ほど好ましい評価であるとする．プロセスの期間 (時間ステップ) が有限でも無限でも，信念状態 (belief state) が確率 $b(s)$ のベクトルとして与えられれば，最適な value function は下に凸型の関数となる [Smallwood et al. 73]．最適な Q-function も同様に下に凸型の関数となる．凸型となる原因については直観的に以下のように説明できる．信念状態空間の中央部付近の領域は，エージェントが現時点の状態認識について区別のつかない状況を示しており，そこではあまり適切な行動選択ができないため value function が下に凸となる．

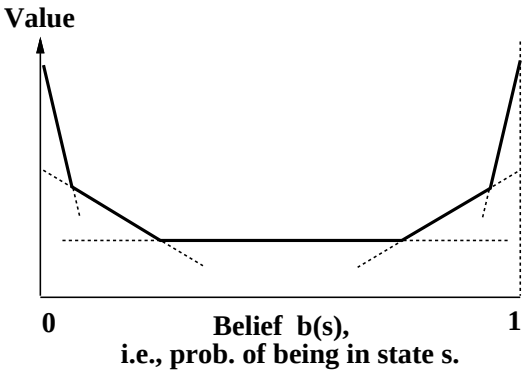



図 2　凸型多面体 (PWLC) の value function．

〔 4 〕 凸型多面体 (**PWLC**) の **Value Function**

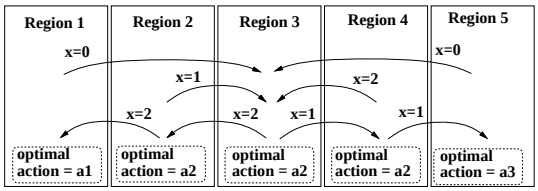
Belief MDP における有限期間の最適な value function は全て凸型の多面体関数 (piecewise linear and convex: PWLC) である (図 2)．無限期間の最適な value function が厳密に凸型の多面体関数となるような POMDP のクラスが存在し，それは "finitely transient" と呼ばれている [Smallwood et al. 73]．また，全ての無限期間の最適な value function は任意の精度で凸型の多面体関数で近似可能である [Sondik 78]．ある Value function $V^\pi_\infty(b)$ が凸型の多面体関数のとき，以下のように表すことができる．

$$V^\pi_\infty(b) = \max_{\alpha \in L_a} \alpha \cdot b \ , \qquad (1)$$

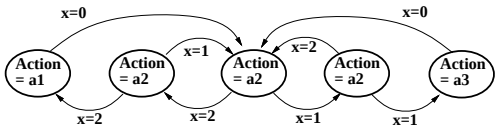
ただし $L_a$ は $|S|$ 次元ベクトルの任意の集合である．最適な Q-function も同様に表せる．

〔 5 〕 信念状態空間の有限離散表現

有限期間の場合または "finitely transient" な場合においては，連続的で不加算無限な信念状態の空間は，$L_a$ で示される超平面によって凸型領域の有限な集合へ分割できる [Smallwood et al. 73]．同一領域中の信念は全て等価である．これにより，状態空間が連続な信念状態の MDP(belief MDP) は，離散状態の MDP




**(a) A state transition diagram of the Markov partition over the belief space.**




**(b) A corresponding plan graph.**

図 3　信念状態空間の有限離散表現の例とプラングラフ．

へと変換できる．トラの例題に関する belief MDP を図 3(a) に示す．

〔 6 〕 プラングラフ

最適な行動とその直後の観測が与えられると，信念状態空間中の分割された領域間の状態遷移は決定的である．領域の集合とその間の状態遷移によってプラングラフ [Kaelbling et al.] が構成できる．これは，最適な政策の行動選択を縮約したものである．プラングラフを用いて行動選択を行う場合，信念状態や value を維持するような計算コストのかかる処理が不要になる．トラの例題に関する belief MDP とそれに対応するプラングラフを図 3 に示す．

### 2・4 POMDP の解法

POMDP の正確なモデルが与えられた場合，DP に基づくいくつかのアルゴリズムが提案されている．Sondik の one-pass アルゴリズムや Cheng の Linear support アルゴリズム [Lovejoy 91]，Witness アルゴリズム [Cassandra et al. 94] は，式 1 に示すような凸型多面体関数を利用して Value function を求める．[Parr et al. 95] では，連続で微分可能な関数を用いて無限期間の Value function を近似する方法 (Smooth Partially Observable Value Approximation: SPOVA) を提案している．

残念ながら厳密な最適 Value function や最適政策を求めることは以下の理由により計算量的に実行不可能である．Value function を表すためのベクトル集合 $L_a$ の要素数が，考慮する期間に対して指数的に増加するからである．また，与えられた POMDP の厳密な最適政策を求める計算量は，たとえ有限期間であっても PSPACE-complete だと言われている [Papadimitriou et al. 87]．よってより高速な近似アルゴリズムについての研究が行われている [Littman et al. 95a]．

## 3. 強化学習アルゴリズム

強化学習の枠組では，エージェントは事前に環境に関する知識を持っていない．すなわち，エージェントは状態遷移確率 $P^a_{ss'}$ や観測の確率 $P(x|s)$ や報酬の確率 $R^a(s)$ についての情報を持っていない．POMDP のモデルを与えた場合に最適政策を得る計算量が PSPACE-complete であり，時系列データから隠れマルコフモデルを学習する問題の計算量は NP-hard である [Abe et al. 92] ことより，これらの複合問題である強化学習問題の複雑さは非常に大きいものとなる．複雑さが非常に小さい問題を除き，厳密解を得ることはほとんど不可能に近い．そのため多くのアルゴリズムは近似解法である．本章ではこれらのアルゴリズムについて概説する．

### 3・1 メモリレスな政策の学習

本節では観測から行動への直接的な写像を政策としたときに，よい政策を得るための強化学習アルゴリズムについて紹介する．これは，ノイズを含んだ観測入力をそのまま状態入力とする単純な接近法だが，アルゴリズムは全て極めて単純になり，学習すべきパラメータ数も非常に小さくてすむので，パフォーマンスの改善が早いなどの利点がある．

〔 1 〕 メモリレスな決定的政策

Q-learning (one-step Q-learning) は動的計画法 (dynamic programming:DP) に基づく逐次的な強化学習アルゴリズムである [Watkins et al. 92]．ここでは，Q 値を状態空間の代わりに観測の空間に対して用いることにする．それぞれの時間ステップで以下のように更新する．

$$\Delta Q = r_t \ + \ \gamma \max_{u \in A} Q(u, X_{t+1}) - Q(a_t, X_t),$$
$$Q(a_t, X_t) \ \leftarrow \ Q(a_t, X_t) + \alpha \, \Delta Q,$$

ただし $\alpha$ は非負の学習係数である．観測の不確実性や不完全性の小さな POMDP の環境ならば，このように Q-learning をそのまま適用してうまくいく場合がある．しかし，多くの場合においてパフォーマンスがかなり悪くなる．

〔 2 〕 メモリレスな確率的政策

POMDP において現在の観測だけから行動を決定

する場合には，決定的な政策よりも確率的な政策を用いた方がよい場合があり，さらにその場合の最適性の評価規範として平均報酬が最も適していることが示されている [Singh et al. 94]．ここで確率的な政策とは，観測から行動出力への確率分布である．平均報酬について局所最適な確率的政策を求める強化学習アルゴリズムとして，モンテカルロ法による政策評価と山登りによる政策改善を組み合わせた方法が提案されている [Jaakkola et al. 94]．モンテカルロ法による政策評価 (Monte-Carlo policy evaluation) は以下の手順で行う．ある確率的政策のもとで行動選択を行い，平均報酬で定義される $Q(X,a)$ 値を計算する．現在の行動として $a_t = a$ を実行したとき，

$$\begin{aligned}
\beta_t(X,a) &= (1 - \frac{1}{k_t(X,a)})\,\gamma_t\,\beta_{t-1}(X,a) \\
&\quad + \frac{1}{k_t(X,a)} \\
Q_t(X,a) &= (1 - \frac{1}{k_t(X,a)})\,Q_{t-1}(X,a) \\
&\quad + \beta_t(X,a)\ (r_t - \bar{r}) \\
\beta_t(X) &= (1 - \frac{1}{k_t(X)})\,\gamma_t\,\beta_{t-1}(X) + \frac{1}{k_t(X)} \\
V_t(X) &= (1 - \frac{1}{k_t(X)})\,V_{t-1}(X) \\
&\quad + \beta_t(X)\ (r_t - \bar{r})
\end{aligned}$$

また，$a_t \neq a$ の Q 値については，

$$\begin{aligned}
\beta_t(X,a) &= \gamma_t\,\beta_{t-1}(X,a) \\
Q_t(X,a) &= Q_{t-1}(X,a) + \beta_t(X,a)\ (r_t - \bar{r}) \\
\beta_t(X) &= \gamma_t\,\beta_{t-1}(X) \\
V_t(X) &= V_{t-1}(X) + \beta_t(X)\ (r_t - \bar{r})\,,
\end{aligned}$$

ただし，$k_t(X,a)$ は観測-行動のペア $(X,a)$ の生じた回数を表し，$k_t(X)$ は観測 $X$ の生じた回数，$\bar{r}$ は $r_t$ の平均値，$\gamma_t$ は最終的には 1 に収束するような割引率である．また，確率的政策の改善は以下のように行う．同じ観測 $X$ の中で $Q(X,a)$ が最大の値を持つ行動の選択確率を増やすことで政策が改善される．観測 $X$ において $\max_a [Q(X,a) - V(X)] > 0$ を満たす限り，平均報酬を増加させる方向へ政策が改善されることが保障される．

上記の Jaakkola らの手法のように明示的に value の推定を行わずに確率的な傾斜法によって確率的政策を改善する方法も提案されている [Williams 92], [Kimura et al. 95]．観測 $X$ においてエージェントが行動 $a$ を選択する確率を，パラメータ $W$ を用いて関数 $\pi(a,W,X)$ で表す (図 4)．パラメータ $W$ はエージェントの内部変数ベクトルを表す．エージェントは内部変数 $W$ を調節することにより確率的政策 $\pi$ を変えることができる．行動選択確率を表す機構が，例えばニューラルネットならば，内部変数 $W$ はリンクの重み変数に相当し，重み付きのルールベースシステムならば，$W$ はルールの重みに相当する．確率的傾斜法による強化学習アルゴリズムの一般形を図 5 に示す．

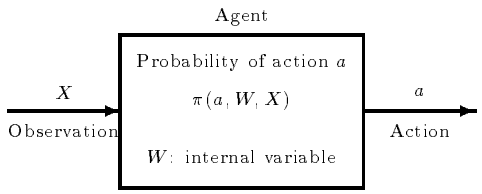



図 4　内部変数パラメータ W を用いて関数表示された確率的政策．

(1)　環境の観測 $X_t$ を受けとる．
(2)　$\pi(a_t, W, X_t)$ の確率で行動 $a_t$ を実行する．
(3)　環境から報酬 $r_t$ を受け取る．
(4)　内部変数 $W$ の全ての要素 $w_i$ について以下の $e_i(t)$ と $D_i(t)$ を求める．ただし $\gamma$ は割引率 $(0 \leq \gamma < 1)$ である．

$$\begin{aligned}
e_i(t) &= \frac{\partial}{\partial w_i} \ln\Big(\pi(a_t, W, X_t)\Big)\,, \\
D_i(t) &= e_i(t) + \gamma D_i(t-1)\,.
\end{aligned}$$

(5)　以下の式を用いて $\Delta w_i(t)$ を求める．

$$\Delta w_i(t) = (r_t - b)\,D_i(t)\,,$$

ただし $b$ は定数である．
(6)　政策の改善: 以下の式で $W$ を更新

$$\begin{aligned}
\Delta W(t) &= (\Delta w_1(t), \Delta w_2(t) \cdots \Delta w_i(t) \cdots)\,, \\
W &\leftarrow W + \alpha(1-\gamma)\,\Delta W(t)\,,
\end{aligned}$$

ただし $\alpha$ は非負の学習定数である．
(7)　時間ステップ $t$ を $t+1$ へ進めて，1 へ戻る．

図 5　確率的傾斜法による強化学習法の一般形

これらの手法は並列処理や政策のパラメータ表現の容易さなどで利点があり，DP に基づいていないという特徴がある．これは，状態観測に未知の不確実性が存在すると最適性原理が成り立たず，DP が使えないためである．

### 3・2　内部状態表現による政策の学習

POMDP の環境において本当に効率的に振舞うためには，信念状態 (belief state) と同等なあるいはプラングラフに相当するような何らかの内部状態表現を用いることが必要である．内部状態表現の違いによって様々なアプローチが存在する．

〔 1 〕 モデルベースによる内部状態表現

予測的区別による接近法 (predictive distinctions

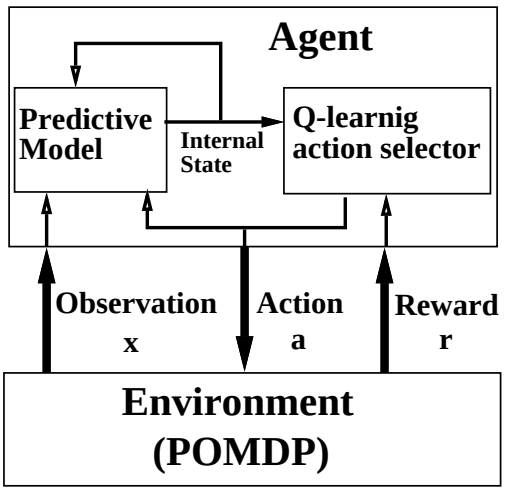



図 6　モデルベースによる内部状態表現．

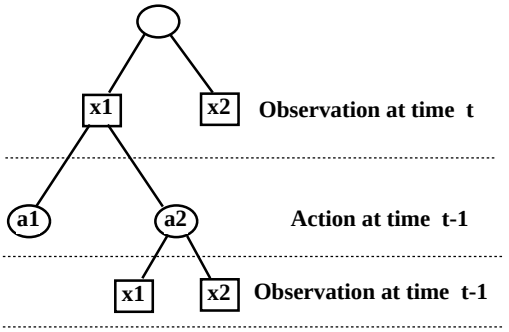



図 7　木構造で表された過去の履歴．ルートからそれぞれの葉ノードに至るパスが可変長の履歴を表す．

approach)[Chrisman 92] や Utile Distinction Memory (UDM) [McCallum 93] では，エージェント内部に環境の POMDP モデルを生成し，信念状態空間の MDP(belief MDP) を解くことにより行動決定を行う (図 6)．まずランダムに生成した小さな POMDP モデルにおいて，エージェントの経験した時系列データを用いてトレーニングを行う．次に，POMDP モデルについて統計的検定を行い，POMDP の状態をさらに増やすかどうかを決める．状態が足りなかったら，状態を分割して同じことを繰り返す．Chrisman の方法と McCallum の方法の違いは，状態が不足しているかどうかを検定する部分である．Chrisman の方法では，政策と状態遷移確率の独立性を検定しているのに対し，McCallum の方法では Q 値の統計的検定を行う．適切な POMDP モデルを生成して belief を計算した後は，**2・4** 節と基本的に同じであるが，ただ 1 つの線形関数で Q 値を近似する簡便なアルゴリズムで対処している．

〔 2 〕 有限長の過去の履歴による内部状態表現

信念状態 (belief state) は，**2・3** 節で述べたように，現在の時点に至るまでの観測と行動の履歴を用いて計算される．従って観測と行動の履歴を有限ステップで打ち切り，そのまま内部状態表現とすれば，信念状態 (belief state) と近似的に同等な表現となることが期待できる．Window-Q architecture [Lin et al. 92], [Whitehead et al. 95] では，長さが固定された状態-行動の履歴を内部表現の状態として Q-learning を行う方法を提案している．一般に，非マルコフ性を解消するのに十分な履歴の長さはエージェントにとって未知である．そのため，履歴の長さを固定すると，不必要に大きな状態空間を指数的大きさで生成したり，長さの不足のため非マルコフ性を解消できない場合がある．さらに，一般に考慮すべき履歴の長さは，信念状態の位置によってまちまちである．例えば，状態観測に不確実性のない状態にいる場合には，過去の経験までさかのぼって考えることは無意味である．

Utile Suffix Memory (USM)[McCallum 95b] では，過去の履歴を木構造で表現し，それぞれの葉ノードを内部状態とすることにより，可変長の履歴 ( 図 7 ) を扱う．それぞれの葉ノードに対する Q 値を Q-learning で学習させている．この Q 値に関して，Kolmogorov-Smirnov 検定を行い，非マルコフ性が存在する場合にはさらに枝を伸ばす．この処理を繰り返すことで，非マルコフ性を排除するのに必要かつ十分な履歴の長さを得るものである．また [Suematsu et al. 97] でも木構造の可変長の履歴を用いて環境モデルを生成している．各葉ノードは，別の葉ノードへの遷移確率を保持しており，ベイズ統計に基づいて事後確率を最大化するモデルを選択する．

〔 3 〕 リカレントネットワークによる内部状態表現

Recurrent-Q (図 8a) では，現時点の状態-行動のペアをリカレントニューラルネットワークへ入力して Q-learning を行う [Lin et al. 92], [Whitehead et al. 95]．各時間ステップにおける観測-行動入力と直前のステップにおける時系列的な特徴を用いれば，リカレントニューラルネットが自動的に新たな時系列的特徴を生成していくだろうというアイデアである．これは、現時点の観測 $x_t$ と 1 ステップ前の信念を用いて逐次的に信念状態を更新するのと類似している．しかし，この接近法では時系列の特徴の学習と Q 値の学習の複合問題を同時に扱うことになるため，学習が非常に困難である．

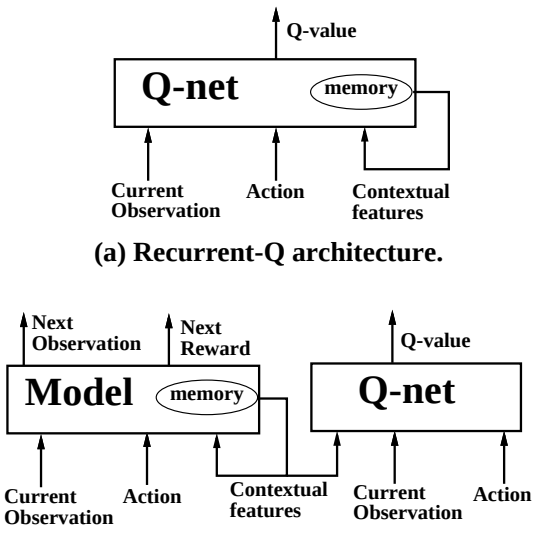



図 8　リカレントニューラルネットによる内部状態表現．

Recurrent-model architecture(図 8b) では，リカレントニューラルネットを用いて次のステップにおける観測の予測を学習すると同時に Q 値を学習する [Lin et al. 92], [Whitehead et al. 95] . この接近法では，履歴の特徴について学習する部分と Q-learning を行う部分が別々になっているので，いくらか効率よく学習できる．履歴の特徴についての学習は，信念状態 (belief state) を計算するのに必要な $P(x|s)$ を学習することと等価であると考えられる．また，履歴の特徴空間における Q-learning は信念状態空間の MDP(belief MDP) における Q-learning と等価であるとも考えられる．よって，リカレントモデルはモデルベースの接近法とも関係する．

〔 4 〕 エキスパートの知識を利用した内部状態表現

大規模な実問題において、何も知識のない初期状態から学習を始めることは事実上実行不能である。環境の状態遷移や状態観測の確率などに関する知識が利用可能な場合には、できるかぎりこれを用いて網羅探索的な試行を極力減らすことが求められる。エージェントが事前に知識として環境の完全な POMDP モデルを持っている場合，学習問題は POMDP 下でのプラニングとなる．

POMDP の環境における強化学習問題において，エキスパートの知識や既存のデータを効果的に利用するアプローチとして，確率ネットワーク (probabilistic network) あるいはベイジアンネットワーク (Bayesian network)[Russell et al. 95] と呼ばれる知識表現を用いた推論を利用する方法が提案されている．このネットワークは，観測データを用いてパラメータを更新していくことにより，事前知識の欠落や誤りを補うことができる．[Forbes et al. 95] では，自動車の自動運転の問題を POMDP における強化学習と考え，確率ネットワークの学習によって信念 (belief) を生成して強化学習を行った．この事前知識を用いる方法もモデルベースの方法と関係する．

〔 5 〕 プラングラフによる内部状態表現

POMDP のモデル推定や Value の推定を一切行わずに，最適な政策の行動選択を要約したプラングラフと等価なものを直接探索する試みがある．Wiering と Schmidhuber は、Levin's search と呼ばれる全探索法の一種を用いてプラングラフと等価なプログラムを得るアルゴリズムを提案し，大規模だが複雑さ (complexity) の低い迷路問題に適用してよい成果をあげている [Wiering et al. 96] . しかし，生成検査法であるため，問題のクラスに制約があり，状態遷移が決定的で単にゴールを目指す程度の複雑さの問題でないと扱えない．

## 4. 考　　察

### 4・1　仮想的な信念状態 (Pseudo-Belief)

Wiering らのアルゴリズムを除いて，内部状態表現を用いるアルゴリズムの多くは何らかの方法で信念 (belief) と等価な機能を持つ仮想的な信念 (pseude-belief) を生成し，それを状態空間とした仮想信念空間の MDP(pseudo-belief MDP) へと問題を変換していると考えられる．仮想的な信念状態が，現在の時点に至るまでのすべての観測と行動の履歴および初期状態分布を利用して行動決定する場合において，十分な情報を含んでいれば，真の信念空間の MDP と等価な問題へと帰着できるだろう．有限長の経験の履歴を利用する方法は精度よく belief space を近似できることが期待されるが，それを保証するためにはさらなる解析が必要である．

### 4・2 「次元の呪縛」問題

連続空間の MDP である belief MDP を離散 MDP に近似するため，信念状態空間の離散的な分割を試みると「次元の呪縛」問題に直面する．信念空間 (belief space) は真の状態数を次元の次数とする連続な超空間であるため，単純に格子状に分割すると状態数が指数

的に増大する．この問題の回避には，従来の強化学習の知見が応用できる．McCallum の USM アルゴリズムは，観測データに基づいて履歴の木構造を生成することで，無駄な内部状態の生成を抑制し，状態空間の爆発を防いでいると考えられる．

### 4・3 内部モデルの学習

Chrisman のアルゴリズムと McCallum の UDM ではモデルのパラメータ更新法として Baum-Welsh アルゴリズムを用いている．これは，ベイジアンネットワーク [Russell et al. 95] やリカレントモデル [Whitehead et al. 95] の更新法と類似している．これらの更新は信念状態 (belief state) を計算するためのパラメータ $P(x|s)$ ををを学習するのと等価であると考えられる．

### 4・4 凸型 Value Function の利用

仮想的な信念を生成するアプローチでは，Value function の凸型の性質を利用する方法は皆無である．その理由は真の信念状態空間と仮想的な信念状態空間の位相空間 (topological space) の違いによるものと考えられる．一般に，Value function の凸型の性質が仮想的な信念状態空間でも成り立つかどうかは不明である．さらに，仮想的な信念を離散表現とする場合が多く，そのとき真の信念のような空間の位相がないことも一因と考えられる．

## 5. お わ り に

本論文では POMDP のいくつかの重要な特徴について簡潔に述べ，いくつかの典型的な強化学習法について概説し，POMDP のどんな性質に依存しているのかについて考察した．POMDP の性質をもっとうまく利用することにより，今後さらに効率的なアルゴリズムが示されることが期待される．

### 謝　　辞

本稿をまとめるにあたり，NEC C & C 研究所の安倍直樹さん，Brown Unv. の Anthony R. Cassandra さん，広島市立大学の末松伸朗さんから重要なコメントをいただきました．ここに感謝の意を表します．

## ◇ 参 考 文 献 ◇

[Abe et al. 92] Abe, N. & Warmuth, M. K.: On the Computational Complexity of Approximating Distributions by Probabilistic Automata, *Machine Learning, 9*,, pp. 205-260 (1992).

[Cassandra et al. 94] Cassandra, A. R. & Kaelbling, L. P. & Littman, M. L.: Acting Optimally in Partially Observable Stochastic Domains, *Proceedings of the 12th National Conference on Artificial Intelligence*, Vol. 2, pp. 1023-1028 (1994).

[Cassandra et al. 96] Cassandra, A. R. & Kaelbling, L. P. & Kurien, J. A.: Acting under Uncertainty: Discrete Bayesian Models for Mobile-Robot Navigation, *Proceedings of the 1996 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 963-972 (1996).

[Chrisman 92] Chrisman, L.: Reinforcement learning with perceptual aliasing: The Perceptual Distinctions Approach, *Proceedings of the 10th National Conference on Artificial Intelligence*, pp. 183-188 (1992).

[Forbes et al. 95] Forbes, J. & Kanazawa, K.& Russell, S.: The BATmobile: Towards a Bayesian Automated Taxi, *14th International Joint Conference on Artificial Intelligence*, pp. 1878-1885 (1995).

[Howard 60] Howard, R. A.: Dynamic Programming and Markov Processes, The MIT Press, Cambridge, Massachusetts (1960).

[Jaakkola et al. 94] Jaakkola, T. & Singh, S. P. & Jordan, M. I.: Reinforcement Learning Algorithm for Partially Observable Markov Decision Problems, *Advances in Neural Information Processing Systems 7*, pp.345-352 (1994).

[Kaelbling et al.] Kaelbling, L. P., & Littman, M. L., & Cassandra, A. R.: Planning and Acting in Partially Observable Stochastic Domains, *Artificial Intelligence*, to appear.

[Kaelbling et al.96] Kaelbling, L. P., & Littman, M. L., & Moore, A. W.: Reinforcement Learning: A Survey, *Journal of Artificial Intelligence Research*, Vol. 4, pp. 237-277 (1996).

[Kimura et al. 95] Kimura, H. & Yamamura, M. & Kobayashi, S.: Reinforcement Learning by Stochastic Hill Climbing on Discounted Reward, *Proceedings of the 12th International Conference on Machine Learning*, pp. 295-303 (1995).

[Lin et al. 92] Lin, L. J. & Mitchell, T. M.: Reinforcement Learning With Hidden States, *Proceedings of the 2nd International Conference on Simulation of Adaptive Behavior*, pp. 271-280 (1992).

[Littman 94b] Littman, M. L.: Memoryless Policies: Theoretical Limitations and Practical Results, *Proceedings of the 3rd International Conference on Simulation of Adaptive Behavior*, pp. 238-245 (1994).

[Littman et al. 95a] Littman, M. L. & Cassandra, A. R. & Kaelbling, L. P.: Learning policies for partially observable environments: Scaling up, *Proceedings of the 12th International Conference on Machine Learning*, pp. 362-370 (1995).

[Littman et al. 95b] Littman, M. L. & Dean, T. L. & Kaelbling, L. P.: On the Complexity of Solving Markov Decision Problems, *Proceedings of the 11th International Conference on Uncertainty in Artificial Intelligence*, (1995).

[Lovejoy 91] Lovejoy, W. S.: A Survey of Algorithmic Methods for Partially Observed Markov Decision Processes, *Annals of Operations Research 28*, pp. 47-65 (1991).

[McCallum 93] McCallum, R. A.: Overcoming Incomplete Perception with Utile Distinction Memory, *Proceedings of the 10th International Conference on Machine Learning*, pp. 190-196 (1993).

[McCallum 95b] McCallum, R. A.: Instance-Based Utile Distinctions for Reinforcement Learning with Hidden State, *Proceedings of the 12th International Conference on Machine Learning*, pp. 387-395 (1995).

[Monahan 82] Monahan, G. E.: A Survey of Partially Observable Markov Decision Processes: Theory, Models, and Algorithms, *Management Science*, Vol. 28, No. 1, January 1982, pp. 1-16.

[Papadimitriou et al. 87] Papadimitriou, C. H. & Tsitsiklis, J. N.: The complexity of Markov decision processes, *Mathematics of Operations Research 12 (3)*, pp. 441-450 (1987).

[Parr et al. 95] Parr, R. & Russell, S.: Approximating Optimal Policies for Partially Observable Stochastic Domains, *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pp. 1088-1094 (1995).

[Rabiner 89] Rabiner, L. R.: A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE 77 (2)*, (1989).

[Russell et al. 95] Russell, S., Binder, J. & Kanazawa, K.: Local learning in probabilistic networks with hidden variables, *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pp. 1146-1152 (1995).

[Simmons et al. 95] Simmons, R., & Koenig, S.: Probabilistic Robot Navigation in Partially Observable Environments, *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pp. 1080-1087 (1995).

[Singh et al. 94] Singh, S. P.& Jaakkola, T. & Jordan, M. I.: Learning Without State-Estimation in Partially Observable Markovian Decision Processes, *Proceedings of the 11th International Conference on Machine Learning*, pp. 284-292 (1994).

[Smallwood et al. 73] Smallwood, R. D. & Sondik, E. J.: The Optimal Control of Partially Observable Markov Processes over a Finite Horizon, *Operations Research 21*, pp. 1071-1088 (1973).

[Sondik 78] Sondik, E. J.: The Optimal Control of Partially Observable Markov Processes Over the Infinite Horizon: Discounted Costs, *Operations Research 26*, pp. 282-304 (1978).

[Suematsu et al. 97] Suematsu, N. & Hayashi, A. & Li, S.: A Bayesian Approach to Model Learning in Non-Markovian Environments, *Proceedings of the 14th International Conference on Machine Learning*, pp. 349-357 (1997).

[Watkins et al. 92] Watkins, C. J. C. H., & Dayan, P.: Technical Note: $Q$-Learning, *Machine Learning 8*, pp. 279-292 (1992).

[Whitehead et al. 95] Whitehead, S. D., & Lin, L. J.: Reinforcement learning of non-Markov decision processes, *Artificial Intelligence 73*, 271-306 (1995).

[Wiering et al. 96] Wiering, M. & Schmidhuber, J.: Solving POMDPs with Levin Search and EIRA, *Proceedings of the 13th International Conference on Machine Learning*, pp. 534-542 (1996).

[Williams 92] Williams, R. J.: Simple Statistical Gradient Following Algorithms for Connectionist Reinforcement Learning, *Machine Learning 8*, pp. 229-256 (1992).

## 著 者 紹 介

木村 元( 正会員 )

1992 年東京工業大学工学部制御工学科卒業．1994 年同大学院総合理工学研究科知能科学専攻修士課程修了．1997 年同専攻博士課程修了．同年 4 月，日本学術振興会特別研究員．同年 5 月，東京工業大学大学院総合理工学研究科 PD 研究員，現在に至る．人工知能，特に強化学習に関する研究を行っている．計測自動制御学会，日本ロボット学会会員．〈gen@fe.dis.titech.ac.jp〉

**Leslie Pack Kaelbling**

Associate Professor of Computer Science Department, Brown University.
〈lpk@cs.brown.edu〉