論　文　　Technical Papers

ロボットア– ムのほふく行動の強化学習：確率的傾斜法による接近

Reinforcement Learning for Crawling Robot Motion Using Stochastic Gradient Ascent

木村　　元[*1]　小林　重信[*2]
Hajime Kimura　Shigenobu Kobayashi

＊1　東京工業大学大学院総合理工学研究科
Graduate School of Interdisciplinary Science and Engineering, Tokyo Institute of Technology

19YY 年 MM 月 DD 日　受理

Keywords: Reinforcement learning, POMDPs, function approximation, stochastic policy, gradient ascent

Summary

Many previous works in reinforcement learning (RL) are limited to Markov decision processes (MDPs). However, a great many real-world applications do not satisfy this assumption. RL tasks of real world can be characterized by two difficulties: Function approximation and hidden state problems. For large and continuous state or action space, the agent has to incorporate some form of generalization. One way to do it is to use general function approximators to represent value functions or control policies. Hidden state problems, which can be represented by partially observable MDPs (POMDPs), arise in the case that the RL agent cannot observe the state of the environment perfectly owing to noisy or insufficient sensors, partial information, etc. We have presented a RL algorithm in POMDPs, that is based on a stochastic gradient ascent. It uses function approximator to represent a stochastic policy, and updates the policy parameters. We apply the algorithm to a robot control problem, and show the features in comparison with Q-learning or Jaakkola's method. The results shows the algorithm is very robust under the conditions that the agent is restricted to computationally very poor resources.

1. は　じ　め　に

従来の多くの強化学習研究では，環境をマルコフ決定過程 (MDP) でモデル化していた．MDP は，報酬に遅れのある環境や，状態遷移に不連続性や不確実性を含むような環境を精度よく近似可能な数理モデルである．代表的な強化学習アルゴリズムとして Q-learning が提案されている [Watkins et.al 92]．しかし，エージェントが環境の状態を完全に観測可能であるという点において非現実的な設定であった．実世界における強化学習では，隠れ状態と関数近似の 2 つの問題の扱いが求められる．

連続で大きな状態や行動の空間を扱う場合には，エージェントは何らかの汎化処理を行う必要がある．例えば，ニューラルネットのような関数近似を用いて Value function や政策を表現するなどの方法がある [Singh et.al 94]．残念ながらこのような汎化は，MDP の環境であっても問題を非マルコフ的にしてしまう原因となる場合がある．すなわち，連続な状態空間において離散化が粗すぎる場合などに生じる [Moore 95]．[Whitehead 95] の不完全知覚 (perceptual aliasing) は，エージェント内部において注目すべきセンサを選択するなどの処理に起因しており，これもエージェント内部における観測入力の汎化処理における同種の問題と考えられる．ここで本論文が問題として注目するのは，上記のように設計者がエージェントに与えた汎化メカニズムが原因で非マルコフ性が引き起こされるとき，強化学習アルゴリズムがその制約の範囲内でいかにして良い政策を見つけるかという点である．

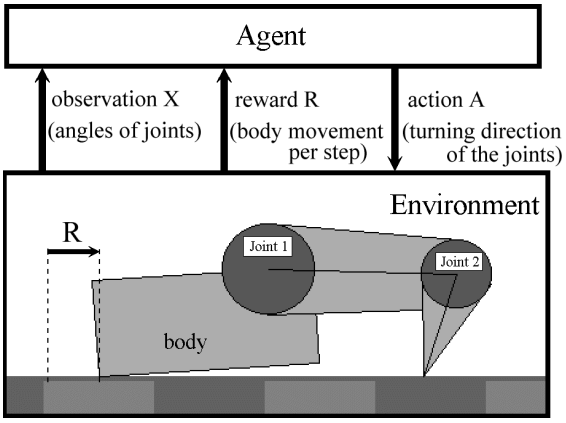


図 1　ロボットアームによるほふく前進行動の学習

一方，隠れ状態問題はエージェントのセンサ能力の制限等により環境の状態観測に不完全性や不確実性を伴う場合に生じ，非マルコフ問題の一種として知られている．隠れ状態問題は POMDP によって記述することができる．本論文ではロボットアームによるほふく行動の強化学習問題を取り上げる．まず，関数近似を用いて POMDP を解く必要が生じることを示す．次に，著者らが提案した確率的傾斜法 [木村 96] を用いて本問題の定式化とシミュレーションを行い，他手法との比較を示し，特徴について考察する．

2. ロボットアームのほふく行動の学習

強化学習の枠組は，故障などやむを得ない状況下で限られたセンサ等を用いて試行錯誤しながら環境に適応するような問題に適すると考えられる．本論文では 2 つの関節を持つロボットアームにボディをひきずって歩くような動作を学習させることを考える (図 1) . 本例題は，状態空間の汎化に起因する不完全知覚問題に注目すれば，多足歩行ロボットの学習と共通の問題点を有していることから，これら多足歩行ロボットの問題に対する基礎的考察を与えると考えられる．本論文では POMDP+関数近似という理論的な問題設定を満たしつつ実世界の問題に接近するため本例題を設定した．

エージェントはある時間間隔で関節の角度を観測し，行動としてそれぞれの関節のモータの回転方向を出力する．次の行動決定までの間に定速モ− タが回り続けて大きなギヤ比で関節を動かし続ける．その結果アームが少し動くので，1 ステップ分のボディの前進距離を報酬として受けとり，次の時間ステップへ進む．この報酬は符号付きで，後退するとその距離に比例した罰が与えられる．ロボットが前進するためには，アーム先端を手前に掻き寄せるだけでなく，アームを持ち上げて前へ出し，また手前に掻き寄せる動作を繰り返さなければならない．環境としてのロボットの状態遷移は，状態空間が連続な POMDP としてモデル化できる．

この問題では以下の 3 点の課題を扱う必要がある．

(課題 1) 報酬の遅れ：　この問題ではアーム先端が図 2の (1) に示す軌跡になるよう学習することが望ましい．アーム先端を A → B → C のように手前に掻き寄せる動作については，ボディが動いて即座に報酬が入るので学習は簡単である．しかし C → D → A のようにアームを持ち上げて前に出す動作は，実行中に報酬が入らず，遅れがあるため学習が困難である．例えば

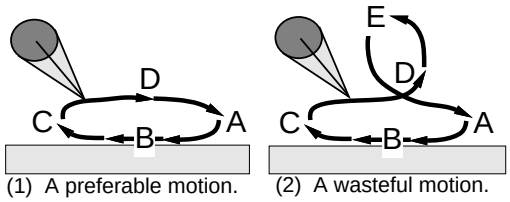


図 2　アーム先端の軌跡の例

エージェントが図 2の (2) のように C → D → E → D → A のように無駄な行動をしても，この間ボディが進まず報酬が入らない点においては正しい行動をとった場合と同じなので，行動を実行した直後では無駄な行動と正しい行動は区別できない．よって無駄な行動を排除して正しい行動を強化するためには，遅れて入ってくる報酬を手がかりにしなければならない．

(課題 2) 隠れ状態問題：　ロボットのセンサ能力等が不十分なため，環境を一部しか観測できなかったり，観測にノイズを伴う場合がある．状態観測の不完全性や不確実性のために隠れ状態問題となる．

(課題 3) 状態空間の汎化に伴う不完全知覚：　強化学習ではエージェントの計算資源は一般に有限なので，連続または大きな状態空間を持つ環境を扱う場合には空間を適切に汎化する必要がある．ニューラルネット等を用いる場合もあるが，マルコフ決定問題では状態空間を多次元のグリッドに分割して離散化するのが一般的である．しかし，マルコフ性を満たすよう十分細かく分割すると，次元数に対する状態の個数は指数関数的に増大する「次元の呪い (curse of dimensionality)」問題が発生する [Moore 95] . 逆に状態空間の分割が粗いと上記の問題を回避できるが，不完全知覚に

よるPOMDP問題を引き起こす．強化学習では，環境はエージェントにとって未知であるため，どの程度まで細かく分割すればよいのかを予め知ることはできない．状態空間の分割が粗ければ，不完全知覚による影響がより強くなる．ここで本論文が問題として注目するのは，上記のように設計者がエージェントに与えた汎化メカニズムが原因で不完全知覚が引き起こされたとき，強化学習アルゴリズムがその制約の範囲内でいかにして良い政策を見つけるかという点である．図3の

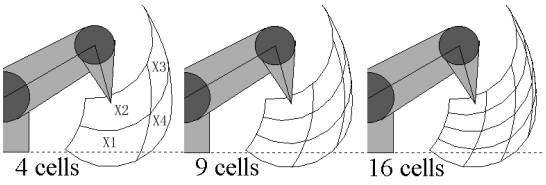


図3　連続値角度センサ空間のグリッド分割による観測の離散化

ように，連続値の関節角度のセンサ入力を4等分してエージェントの観測を離散化することを考える．観測$X1$の領域では，アーム先端が地面に接触している状態と接触していない状態が共存するが，両方とも同じ領域$X1$として観測され，それぞれを区別できない不完全知覚を生ずる．このときエージェントから見た問題は観測入力が4種類しかない困難なPOMDPになる．エージェントはこのように限られた制約の範囲内で前に進むための制御規則を見つけなければならない．

従来の多くの強化学習研究では，観測入力の汎化処理における関数近似能力不足による非マルコフ性の問題は，関数近似能力を十分に持たせることで対処してきた．エージェントにとって環境は未知であるという強化学習の枠組に沿って考えると，エージェントが十分な関数近似能力を持つとは限らない．エージェント内部の汎化処理に起因する「不完全知覚」は，環境そのものに起因する「隠れ状態問題」と数学的にほぼ等価な非マルコフ問題と考えられる．例えば図3で示した不完全知覚は連続値の角度入力を離散化したために生じるが，ロボットのセンサ入力が，最初から離散値の場合は隠れ状態問題になる．どちらの場合も，離散観測を用いて学習する強化学習アルゴリズムからみれば，同じPOMDP環境下での最適化問題である．本論文では，従来見落とされがちな「あらかじめ与えられ，時には不十分かもしれない有限な計算資源」という制約条件を強調するために，環境そのものに起因する「隠れ状態問題」とエージェント内部の汎化処理に起因する「不完全知覚」を区別して示した．

本節で掲げた3つの課題は実世界の問題を取り扱う上で常に要求されることであり，本例題への具体的接近法を示し，その性能評価を行うことは，強化学習の工学的応用の道を拓く上で有用と思われる．

3. 確率的傾斜法による接近

3・1 従 来 手 法

従来，POMDP に対しては，何らかの状態推定器と Q-learning を組合わせるアプローチが主流であった [Chrisman 92], [McCallum 95], [Lin 92], [Whitehead 95] .すなわち，状態推定器では観測系列から現在の真の状態を推定することでPOMDPをMDPに変換する．このMDPの下において，メモリレスな強化学習法であるQ-learningを用いて適切な行動を獲得するものであった．しかしエージェントの持つ限られた計算資源の範囲で状態推定を完全に保証することは一般に困難である．そのため未知なる環境へエージェントを適用する場合には，問題が依然POMDPのままになることを想定しておかなければならない．上記の従来手法の欠点を克服するためには，POMDPにおいてmemory-lessな確率的政策を求める強化学習法が有用と考えられる．

また，状態推定器はしばしば非常に大きなあるいは連続な状態空間を生成しなければならない．強化学習器において何らかの汎化処理と組み合わせることを求められる場合がある [Lin 92] . よってQ-learningに代わり，POMDPの環境で関数近似を用いて学習するようなメモリレスな強化学習が求められている．

離散 POMDPs 下において確率的政策を効率よく求めるための強化学習法として，モンテカルロ法による政策評価と政策改善法を組み合わせた方法がある [Jaakkola 94] . しかしこれは訪れた観測や実行した行動の回数をカウントする必要があり，一般的な関数近似との組み合わせは困難である．

本論文では，観測空間の汎化方法を含む関数近似による政策表現は，最適化問題として与えられた制約そのものであり，その制約の範囲内で性能を最大化できるような強化学習アルゴリズムを追求するという立場をとる．隠れ状態問題および不完全知覚という異なる原因から生じる2つの問題を，関数近似を用いてPOMDPを解く問題と考えることにより統一的に扱う．著者らは，離散時間の有限なPOMDPの環境のもとで，単純に「観測入力に対する行動出力の確率」の関数を最適化するメモリレスな強化学習法として確率的傾斜法 [木村 96] を提案している．これはPOMDPの環境下

で関数近似を用いて確率的政策を学習することが可能である．本論文では，確率的傾斜法を用いて本問題に適用することを考える．

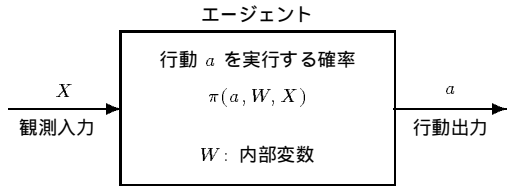


図 4　確率的政策：ベクトル W を調節することで政策を変える．

3・2　確率的政策の関数近似

エージェントの学習目標は，報酬獲得を最大化するように，それぞれの観測において行動を選択する確率分布を形成することである．観測 X においてエージェントが行動 a を選択する確率を政策πと呼び，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πを変えることができる．例えば，行動選択確率を表す機構がニューラルネットで構成されているならば，内部変数 W はリンクの重み変数に相当する．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．$\pi(a, W, X)$ の関数形については，エージェントに実装できる計算資源の制限など，一般に個別の問題ごとに制約が存在する．すなわちエージェントの構造および制約条件は$\pi(a, W, X)$ の関数形で規定される．W をパラメータとした単純な関数として政策 π を記述することの利点は，上記のように様々な構造を持つエージェント全てに対して共通の理論的基礎を与えられることである．

3・3　確率的傾斜法の一般形 [木村 96]

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．ただし全ての観測 X と行動 a において政策 $\pi(a, W, X)$ はベクトル W の全ての要素で偏微分可能であると仮定する．図 5のアルゴリズムでは，W の任意の i 番目の要素を w_i と表す．アルゴリズム中の手順 4 の $e_i(t)$ は適正度 (eligibility) と呼ばれ，実行した行動の情報論的な意味での価値を表す量である．$D_i(t)$ は適正度の履歴(eligibility trace)と呼ばれ，今までとった行動の履歴を記憶しているが，過去の行動ほど割引率 γ で減衰 / 忘却している．エージェントは報酬を受けとると，手順 5,6 にて履歴にある行動の確率を高めるように W を更新する．このような処理を繰り返すと，報酬獲得に関係ない行動は打ち消され，報酬獲得に関係する行動だけが強化される．行動の履歴を強化するので，報酬の獲得に遅れのある行動も強化される．本手法は報酬を受けとった時点で，今までの経験を強化する方法なので経験強化型と呼ばれる学習アルゴリズムに分類される．本手法は割引報酬の期待値を最大化する方向へと政策を確率的に逐次改善するものであり，そこに名前の由来があるが，理論の詳細については [Kimura 95], [木村 96] を参照されたい．

(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γは割引率 $(0 \leq \gamma < 1)$ である．

$$e_i(t) = \frac{\partial}{\partial w_i} \ln\Big(\pi(a_t, W, X_t)\Big) ,$$

$$D_i(t) = e_i(t) + \gamma D_i(t-1) ,$$

(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 は報酬基底と呼ばれる定数である．
(6)　政策の改善: 以下の式で W を更新

$$\Delta W(t) = (\Delta w_1(t), \Delta w_2(t) \cdots \Delta w_i(t) \cdots),$$

$$W \leftarrow W + \alpha(1-\gamma)\, \Delta W(t) ,$$

ただしα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 [木村 96]

4.　実　　験

観測入力の汎化方法や計算資源が限られるために不完全知覚が生じるが，その制約の範囲内で政策をさがすようなロボットのタスクを示す．このタスクのもとで，いくつかの強化学習アルゴリズムの特徴について比較を示す．なるべく公正な比較となるように，アルゴリズムに要する有限な計算資源はほぼ同じとなるように配慮する．

4・1　ロボットの設定の詳細

図 1のロボットの詳細設定を以下に記す．上腕の長さは 34，ひじから先の腕の長さ 20，腕の付け根はボディの左下の隅から水平方向に 32，高さ 18 の所にある．上腕は水平線に対して上方向に 35 度，下方向に 4 度まで回転する．ひじから先の腕は，上腕の軸線に対して上に 10 度，下に 120 度まで回転する．肩の関節のモータは，正 / 逆回転の指令を受けるとその方向へ 12 ± 4 度だけ回転する．ただし分布は一様とする．ひじの関節のモータは，同様に 12 ± 4 度だけ回転す

る．動作限界において，動作範囲を超える方向へ動くような指令を受けた場合，動作限界まで動くがそれ以上は動かないで指令を無視する．アーム先端が地面に接触したら，ボディは滑って動くが，アーム先端は滑ることなく地面をとらえるものとする．アーム先端の滑べりがない点は実ロボットのモデルとして不自然だが，滑べりは報酬入力への単なるノイズと考えられる．実験で用いるアルゴリズムは全て不確実な報酬も扱いが可能であることから，アーム先端の滑べりまで模擬する必要はないと判断した．

4・2　エージェントの実装

(1) 角度センサ入力空間をグリッドで分割/離散化する場合：　図3のように連続値の関節角度のセンサ入力空間を等分割して離散化する．観測 Xは入力ユニット $x_1,\ x_2,\ x_3,\ x_4$ にベクトルとして入力される(図6)．観測ベクトルは，アーム先端の存在する領域に対応する要素だけ 1 で，それ以外は 0 である．例えば観測が X_1のとき，$(x_1, x_2, x_3, x_4) = (1, 0, 0, 0)$ となる．

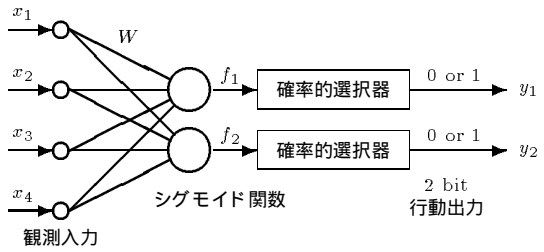


図6　離散値の観測入力の場合の確率的傾斜法を実装するエー ジェントの構造

(2) 連続値の観測(角度)を入力する場合：　各関節の角度を $0 \le \theta_i \le 1$ に正規化して入力する(図7)．

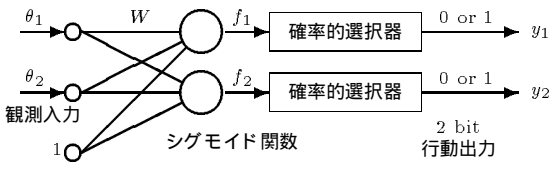


図7　連続値の観測入力の場合の確率的傾斜法を実装するエー ジェントの構造

本章では任意の i 番めの入力ユニットと任意の j番めのシグモイド関数を結ぶ内部変数を w_{ij}と表す．これに伴い，適正度を $e_{ij}(t)$，適正度の履歴を $D_{ij}(t)$ と表す．エージェントは入力ユニット x_iと内部変数 w_{ij} の加重和をとり，評価値 f_jを以下のように計算する．

表1　行動選択確率と適正度

action a_j $= (y_1, y_2)$	action probability $\pi(a_j, W, X)$	eligibility $e(t) = \frac{\partial}{\partial w} \ln \pi(a_j, W, X)$
$a_1 = (0,0)$	$(1-f_1)(1-f_2)$	$\frac{-1}{1-f_1}\frac{\partial f_1}{\partial w} + \frac{-1}{1-f_2}\frac{\partial f_2}{\partial w}$
$a_2 = (0,1)$	$(1-f_1)\,f_2$	$\frac{-1}{1-f_1}\frac{\partial f_1}{\partial w} + \frac{1}{f_2}\frac{\partial f_2}{\partial w}$
$a_3 = (1,0)$	$f_1\,(1-f_2)$	$\frac{1}{f_1}\frac{\partial f_1}{\partial w} + \frac{-1}{1-f_2}\frac{\partial f_2}{\partial w}$
$a_4 = (1,1)$	$f_1\,f_2$	$\frac{1}{f_1}\frac{\partial f_1}{\partial w} + \frac{1}{f_2}\frac{\partial f_2}{\partial w}$

(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_{ij}について以下の $e_{ij}(t)$ と $D_{ij}(t)$ を求める．ただしγは割引率 $(0 \le \gamma < 1)$ である．

$$e_{ij}(t) \;=\; x_i(y_j - f_j)\ ,$$
$$D_{ij}(t) \;=\; e_{ij}(t) + \gamma D_{ij}(t-1)\ ,$$

(5)　以下の式を用いて$\Delta w_{ij}(t)$ を求める．

$$\Delta w_{ij}(t) \;=\; (r_t - b)\, D_{ij}(t)\ ,$$

ただし b は報酬基底と呼ばれる定数である．
(6)　政策の改善: 以下の式で全ての w_{ij}を更新

$$w_{ij} \;\leftarrow\; w_{ij} + \alpha(1-\gamma)\,\Delta w_{ij}(t)\ ,$$

ただしα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8　エージェントに実装した確率的傾斜法

$$f_j = \frac{1}{1 + \exp\left(-\sum_{i=1}^{4} x_i\, w_{ij}\right)}\ . \qquad (1)$$

エージェントの出力において，$y_1,\ y_2$ を 0 または 1 の値をとる確率変数とおく．$f_1,\ f_2$ の値をそれぞれ y_1, y_2が 1 を出力する確率とする．y_1, y_2の出力をそれぞれアームの関節の回転方向に割り当て，確率的に行動を選択する．政策πと適正度は表 1のように表されるが，実際の計算では $f_1,\ f_2$および $y_1,\ y_2$に注目する．エージェントが時間 t で行動として任意の $f_1,\ f_2$を出力したとき，適正度は以下のようになる．

$$e_{ij}(t) \;=\; \begin{cases} \dfrac{-1}{1-f_j}\dfrac{\partial f_j}{\partial w_{ij}} & ,\ \text{where } y_j = 0, \\ \dfrac{1}{f_j}\dfrac{\partial f_j}{\partial w_{ij}} & ,\ \text{where } y_j = 1. \end{cases}$$
$$\;=\; \frac{y_j - f_j}{f_j(1-f_j)}\frac{\partial f_j}{\partial w_{ij}}$$
$$\;=\; x_i(y_j - f_j)\ . \qquad (2)$$

よって図 6, 図 7のエージェントに確率的傾斜法を適用すると，行動選択確率と適正度共にそれぞれのシグモイド関数ユニット fjのローカルな情報だけを用いて計算可能になり，並列分散処理が可能である．式 (2) を図 5の一般形アルゴリズムにあてはめると，図 8のようになる．

エージェントのパラメータは，学習係数 $\alpha = 0.4$, 割引率 $\gamma = 0.95$, 報酬基底 $b = 0.01$ に設定し，重み初

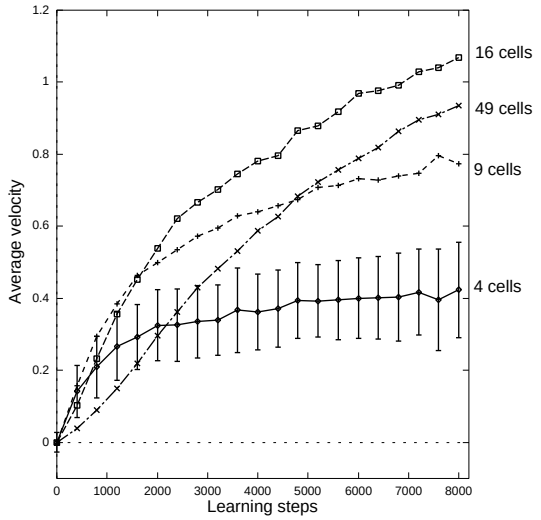


図 9　センサ (角度) 空間をグリッドで離散化入力した場合の 100 試行の学習結果．

期値は ± 0.05 の範囲内でランダムに初期化する．

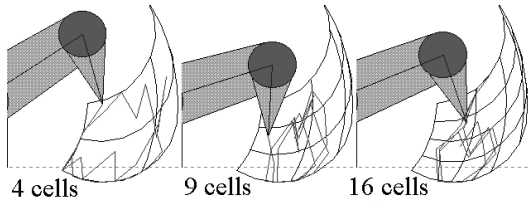


図 10　センサ (角度) 空間をグリッドで離散化入力した場合の学習で得たアーム先端の軌跡の一例．

4・3　実　験　結　果

(1) センサ (角度) 空間をグリッドで分割した場合：図 9 は 4, 9, 16, 49 分割した場合の 100 試行の平均を示す．グラフの横軸は学習ステップ数，縦軸はその学習ステップにおいて政策を固定し，その政策で (学習なしに) 1000 ステップ実行して前進した距離を 1000 で除した値，すなわちその政策の 1 ステップあたりの平均移動距離である．状態空間の分割がたった 4 分割 (4 cells) でも，ある程度の性能を得ている．これは，隠れ状態の存在する観測 $X1$ の領域ではアーム先端を図 10 の左側のように確率的に動かして前進するという政策を発見したことによる．分割数が 9, 16 と大きくなるほど隠れ状態が少なくなり，図 10 の右側のようにきめ細かな制御ができるため，高い平均速度の政策を得ている．しかし分割数が多くなり過ぎると，状態空間が不必要に増大し学習初期の学習速度の低下を招く．

観測入力が離散値に分割されている場合，Q-learning [Watkins et.al 92] や Jaakkola の手法 [Jaakkola 94] との比較が可能である．図 11 から図 16 はそれぞれの状態分割において上記の手法と比較した結果を示している．それぞれのアルゴリズムは 16 分割の問題においてほぼ最適なパラメータの調節を行い，パラメータを固定して全ての問題に適用した．それぞれのアルゴリズムで用いたパラメータ等を以下に記す．Q-learning ではあまり割引率 γ を 1 に近付けると極端に収束が遅くなるため $\gamma = 0.9$ に設定し，学習定数は 0.4 とした．行動選択は最も一般的なボルツマン分布に基づく方法 [Barto et al. 95] を用いた．温度スケジューリングは $temp = \frac{1000}{100 + step}$ として $\exp(Q/temp)$ の比に応じて行動選択を行った．

Jaakkola の方法では $\gamma = 0.95$ に設定し，行動選択確率はそれぞれの観測-行動ペアに定義された重みの比で表した．重みの初期値は 10 に設定し，行動選択確率の更新では最大の Q 値を持つ行動の重みに 0.2 を加えてから重みの合計が同じ値 (40) になるように正規化した．

状態分割が粗く，不完全知覚の度合いが強い場合，図 11, 図 12 のように Q-learning では学習が困難なのに対して Jaakkola の方法と本手法はほぼ同等の性能を得ている．状態分割が細かくなって非マルコフ性が小さくなるにつれて Q-learning の性能が上がっていく (図 16) . 逆に Jaakkola の方法は状態空間の増加に伴い性能が相対的に低下してくる．確率的傾斜法は分割数の変化に対して最もロバストであるが，Jaakkola の方法とはそれほど大きな差はない．ここで強調したいのは，提案手法は，観測− 行動ペアの事象の生起した回数を明示的にカウントする Jaakkola らの手法とは全く異なるアルゴリズムを用いて確率的行動選択を行っているのにもかかわらず，同等の性能を得られただけでなく，次節で述べるように Jaakkola らの手法が適用できないような，任意の関数近似を用いた連続値の観測入力の問題へも適用できる点である．

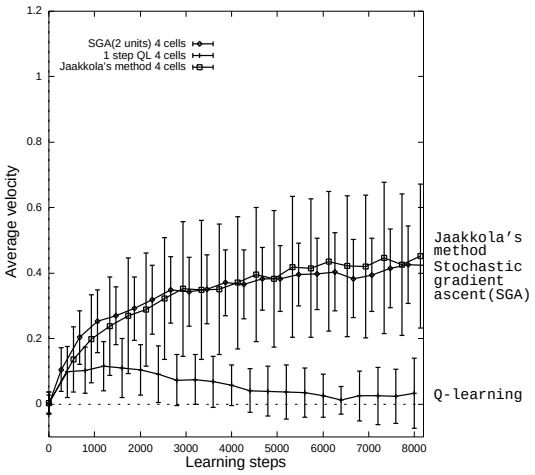


図 11　角度センサの入力空間を 4 分割して離散化した場合の比較

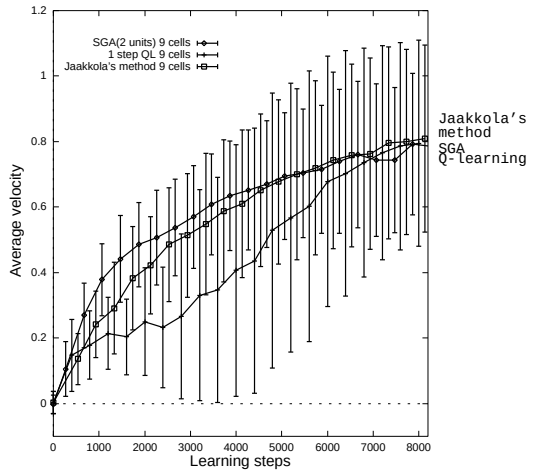


図 12　角度センサの入力空間を 9 分割した場合の比較

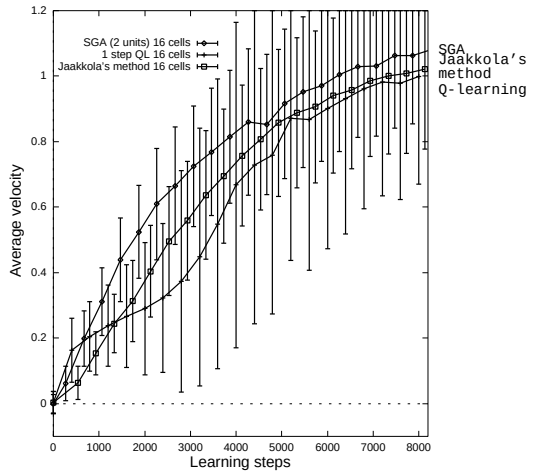


図 13　角度センサの入力空間を 16 分割した場合の比較

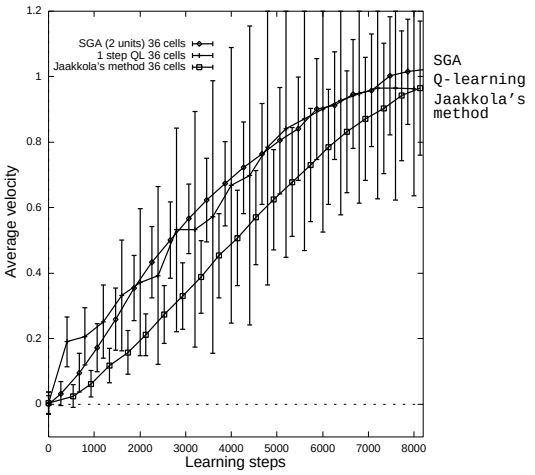


図 14　角度センサの入力空間を 36 分割した場合の比較

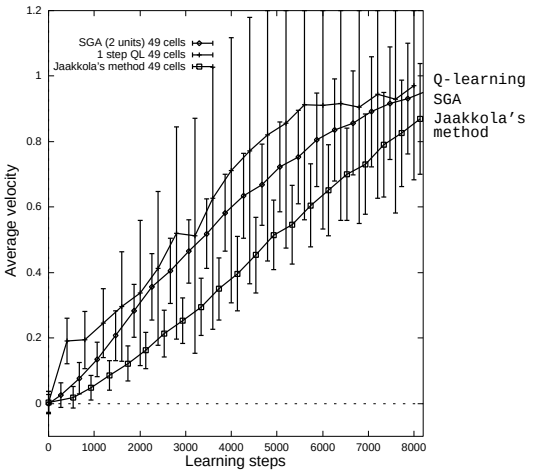


図 15　角度センサの入力空間を 49 分割した場合の比較

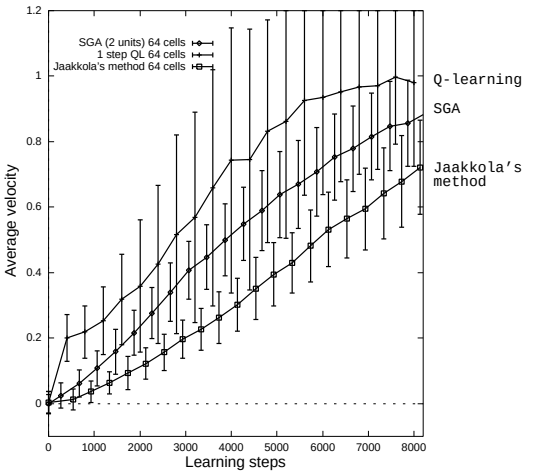


図 16　角度センサの入力空間を 64 分割した場合の比較

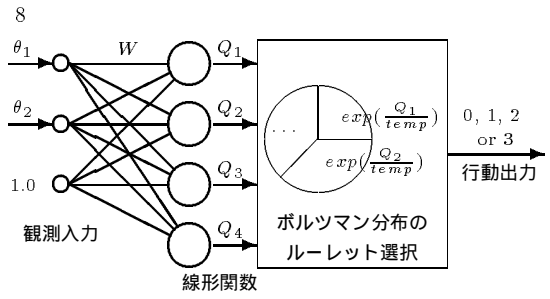


図 17　連続値の観測入力の場合の Q-learning エージェントの実装

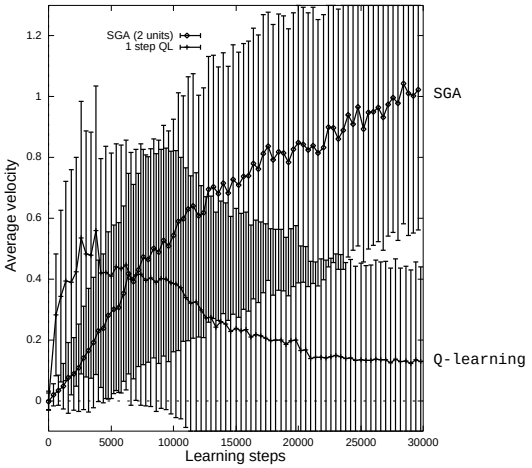


図 18　連続値の観測 (角度) を入力した場合の 100 試行の学習結果

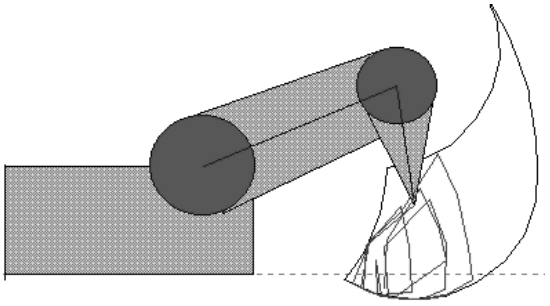


図 19　連続値の観測 (角度) を直接入力して 30,000 ステップ学習後に得たアーム先端の軌跡の一例

(2)連続値の観測(角度)を入力した場合: この実験では , Q-learning との比較を行った . 図 17 に Q-learning の実装を示す . 図 7に示した確率的傾斜法のエー ジェントと異なる構造をとる理由は , 内部変数パラメー タの総数すなわち計算資源をなるべく同じにするためである . 確率的傾斜法は内部変数パラメー タ w_{ij} と同じ個数の適正度の履歴変数 D_{ij} を必要とする . よって図 7の構造に適用された確率的傾斜法の内部変数パラメータの個数と図 17に適用された Q-learning の内部変数パラメー タの個数は同じになるので , 公正な比較が期待できる . また , Q 値の更新は以下の式で行った .

$$\Delta w = \frac{\partial Q(a_t, x_t)}{\partial w} \times \left(r_t + \gamma \left(\max_{u \in A} Q(u, x_{t+1}) \right) - Q(a_t, x_t) \right)$$

Q-learning のパラメータは離散的に分割して入力した場合と同じ値を用い , $\gamma = 0.9$, $\alpha = 0.4$, $temp = \frac{1000}{100 + step}$ に設定した .

図 18は 100 試行の結果の平均を示す . 学習に要するステップ数は , 連続なセンサ入力空間をグリッドで離散化し観測入力とした場合に比べて 2 倍以上になる . Q-learning では学習が困難であることがわかる . Q-learning ではさらに温度冷却をゆっくり行うように $temp = \frac{10000}{1000 + step}$ として実験を行ったが , 性能は下がるという結果を得た . これは , 図 17 に示されたエージェントでは関数近似能力が不十分なため , この問題の value function を適切に関数近似できないことが原因と考えられる .

図 19は 30,000 ステップ学習後に得たアームの動作例を示す . 確率的傾斜法では , 学習が進行しても常に多少の確率的な動きが残った .

5. 考　　察

頑健性: Q-learning では , 状態空間の分割が粗過ぎて非マルコフ性の度合いが強い 4 分割の場合や , 連続値のセンサ入力を用いた場合では学習できなかった . Jaakkola の手法では , 状態空間の分割が粗い場合でも確率的政策を学習したが , 状態分割数が増した場合の学習速度の低下は最も著しく , また , 連続値のセンサ入力を扱うことはできない . これらに対して確率的傾斜法は全ての場合について学習でき , また性能的にも常に上位となっている . 本手法は , 与えられた関数近似能力の範囲で性能を最大化するように学習し , 最も高い頑健性を有している . Q-learning+ボルツマン分布のルーレットによる行動選択は , マルコフ環境において (確率的でなく決定的な) 最適政策を求めるには有効だが , 非マルコフ環境において確率的な政策を得たり , 関数近似能力が不十分な場合において政策を得たりするには向いていないことが実験より分かる . Jaakkola らの手法は , 全ての観測入力-行動出力ペアに対するテーブルが必要になるので , 連続値の状態入力には適用できない . 提案手法は非マルコフな環境においても任意の関数で表された確率的政策を改善していくことが可能

であるという意味において，上記2つの手法と比較して高い頑健性を示すと考えられる．

探査戦略の合理性：訪問回数の少ない領域では訪問回数の多い領域よりも試行錯誤的な経験の回数が少ないのだから，統計的推定論の考え方に従うと，相対的に経験を得るための試行錯誤的な行動選択を多くとるべきである．本手法は，訪れる頻度の多い状態ほど優先的に学習が進む傾向を持っている．一方，頻度の低いところでは相対的にランダムに行動する傾向がある．すなわち，本手法は exploration と exploitation のバランスを考慮した枠組になっている．これは，状態の訪問頻度によって自動的に探索のランダムさを調節するという合理的な探査戦略の機能を含んでいると考えることができる．

エージェントの構造の経済性：本手法のエージェントの構造と制約条件は関数 $\pi(a, W, X)$ で表され，W で微分可能という条件を満たせばよく，しかも極めて限られた計算資源のもとで実行可能である．実験で用いたエージェントでは，行動4コに対してわずか重み変数6コ，シグモイド関数2コ相当の計算資源で学習が可能である．本手法はエージェントの内部構造として，実験で示した構造だけでなく，ニューラルネットやファジィなど様々な関数近似の枠組を適用可能である．

6. おわりに

本論文はロボティクスにおける強化学習問題を取り上げた．センサ能力の不足や計算資源の制約等により環境が POMDP となる場合があることを示し，POMDP の環境において関数近似を用いる強化学習の必要性について述べた．ロボットの問題へ確率的傾斜法を適用し，Q-learning や Jaakkola の手法との比較実験を行った．実験結果より，本手法は与えられた関数近似能力の範囲で性能を最大化するよう学習し，他手法と比較して最も頑健性に優れた．これより，本手法はさらに多様な構造のエージェントへも適用可能であることが期待できる．

本論文では POMDP への対処方法として確率的政策だけを用いているが，限られた計算資源の範囲内で環境の状態推定やモデル推定が可能ならば，それを用いた方がよいことは明らかである．**3・1** 節で述べたように，環境の状態推定やモデル推定を行う従来手法では，環境の大きさがエージェントの限られた計算資源の範囲を越える場合に問題が生じる．今後は，限られた計算資源の範囲内でできる限り環境の状態推定やモデル推定を行い，それを越える部分については確率的傾斜法を用いて確率的政策で対処するという方法が有望ではないかと考えている．

◇ 参 考 文 献 ◇

[Barto et al. 95] Barto, A. G., Bradtke, S. J. and Singh, S. P.: Learning to act using real-time dynamic programming, *Artificial Intelligence 72 (1995)*, 81-138.

[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).

[Jaakkola 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).

[Kimura 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).

[木村 96] 木村 元，山村 雅幸，小林 重信: 部分観測マルコフ決定過程下での強化学習：確率的傾斜法による接近，人工知能学会誌, Vol.11, No.5, pp.761-768 (1996).

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

[Littman 95] 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).

[McCallum 95] 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).

[Moore 95] Moore A. W. & Atkeson, C. G.: The Partigame Algorithm for Variable Resolution Reinforcement Learning in Multidimensional State-spaces, *Machine Learning 21*, pp. 199-233 (1995).

[Singh 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).

[Singh et.al 94] Singh, S. P. and Yee, R. C.: An Upper Bound on the Loss from Approximate Optimal-Value Functions, *Machine Learning, 16*, pp. 227-233 (1994).

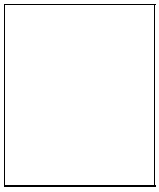
[Watkins et.al 92] Watkins, C. J. C. H., & Dayan, P.: Technical Note: *Q*-Learning, *Machine Learning 8*, pp. 55-68 (1992).

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

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

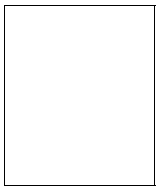
〔担当編集委員：× ○，査読者：○○ ××〕

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年同大学大学院博士課程修了，同年4月日本学術振興会PD研究員，1998年4月，東京工業大学大学院総合理工学研究科助手，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，日本ロボット学会各会員．
〈gen@fe.dis.titech.ac.jp〉



小林　重信（正会員）

1974年東京工業大学大学院博士課程経営工学専攻修了．同年4月，同大学工学部制御工学科助手．1981年8月，同大学大学院総合理工学研究科助教授．1990年8月，教授．現在に至る．問題解決と推論制御，知識獲得と学習などの研究に従事．計測自動制御学会，情報処理学会各会員．
〈kobayasi@dis.titech.ac.jp〉