ロボットアームによるほふく行動の強化学習

○ 木村 元，宮崎 和光，小林 重信
東京工業大学 大学院総合理工学研究科

Reinforcement Learning for Crawling on a Robot Arm

○ Hajime Kimura， Kazuteru Miyazaki， Shigenobu Kobayashi
Interdisciplinary Graduate School of Science and Eng., Tokyo Institute of Technology

Abstract: In this paper, we apply reinforcement learning (RL) to a robot control problem. In general, RL agents are restricted to incomplete sensor readings and finite computational resources. For this reason, the learning problems are considered to be partially observable Markov decision processes (POMDPs). We have presented a RL algorithm for POMDPs, that is based on a stochastic gradient ascent. In this paper, we consider a robot control problem, showing that it is translated as a POMDP. We implement the algorithm to the robot, and we show the features by simulation. The algorithm is very robust under the conditions that the agent is restricted to computationally poor resources.

1 はじめに

自然界における生体の脳は，未知なる環境において，報酬を得て罰から逃れるような適切な行動を，試行錯誤によって獲得するという基本的な適応能力を有している．強化学習は，このシステムを工学的に模倣した枠組である (図 1)．強化学習は，状態の評価に遅れが存在し，かつ状態遷移に離散的な不連続性や不確実性を含むような系における適応的な制御の一種として注目を集めつつあり，幅広い工学的応用が期待される．

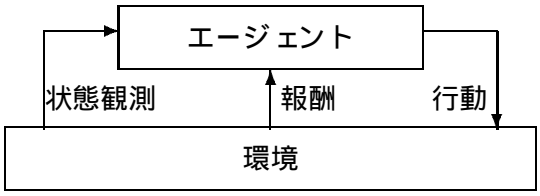

図 1: 強化学習の枠組

例えば，道路の真ん中でロボットの足が壊れたため，とにかく腕などを利用して道路外へ移動しなければならない状況を考える．このように予測困難な突発的状況では，十分に対応可能な制御プログラムが用意されているとは限らない．よって未知なる環境では，ランダムに動いてみて前へ進む方法を模索するというように，試行錯誤で問題解決することが必要だと考えられる．

強化学習では，生体において脳に相当する部分をエージェントと呼ぶ．エージェントは環境の状態を観測して行動を出力するというサイクルを繰り返す．環境はエージェントの行動出力に応じて状態遷移する．このとき，行動と状態に依存して，報酬が環境からエージェントへ与えられる．エージェントの学習目標は，報酬の獲得を最大化するように，観測から行動への写像を形成することである．エージェントは環境に関する知識をあらかじめ持っていないため，試行錯誤を繰り返しながら，より多くの報酬を獲得するように学習する．しかし報酬は行動に対して即座に与えられるとは限らず，遅れを伴うのが一般的なので，どの行動が報酬獲得に貢献したのかを行動の履歴から判断する処理が要求される．

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．MDP の環境においてエージェントの状態観測に不完全性や不確実性があるとき 「隠れ状態問題」と呼ばれ，強化学習における困難な問題として知られるが、これは部分観測マルコフ決定過程 (POMDP) としてモデル化できる [Jaakkola 94] [Singh 94]．POMDP は 「未

知なる環境」という現実に即した自然な設定であるが，従来の Q-learning では学習できない．またエージェントの計算資源は一般に有限なので，実問題のように連続で大きな状態空間の環境を扱う場合には観測入力を汎化するため何らかの関数近似を用いる必要があるが、汎化が不適切だと隠れ状態と同様の非マルコフ問題を生じる．著者らは POMDP の環境において関数近似を用いる強化学習法として確率的傾斜法を提案した [木村 96]．

本論文ではロボティクスにおける強化学習問題を取り上げ，環境が POMDP となることを示す．次に確率的傾斜法を用いたロボットについて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運動の学習

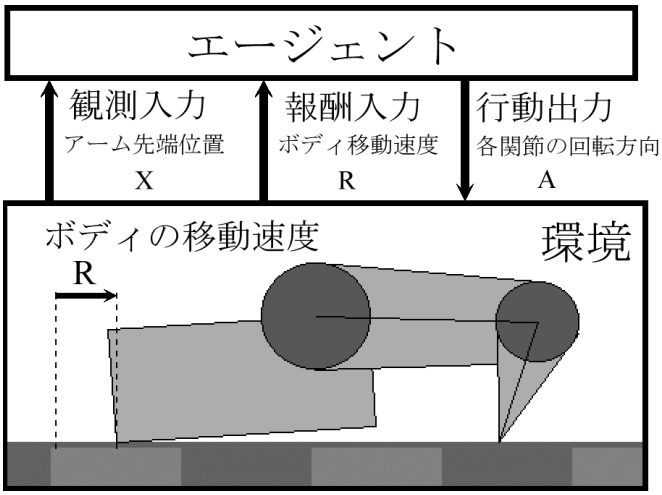

図 2: ロボットアームによるほふく運動

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 (図 2)．エージェントはある時間間隔でアーム先端の位置を観測し，行動としてそれぞれの関節のモーターの回転方向を出力する．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進んだ距離を報酬として受けとり，次の時間ステップへ進む．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アームの状態遷移は，連続な状態空間における MDP 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 ： この学習問題ではアーム先端が図 3の右側のような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．例えばエージェントが図 3の左側のように C

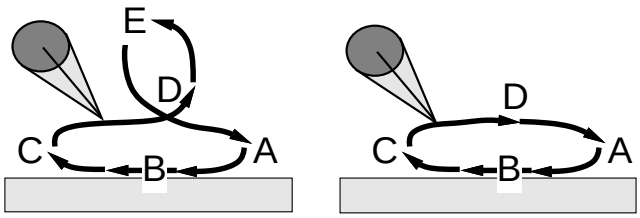

図 3: アーム先端の軌跡の例．

→ 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) 状態空間の汎化と不完全知覚 ： 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．結果的に状態空間の分割が粗くなって POMDP

問題になる．このロボットの問題も状態空間の分割

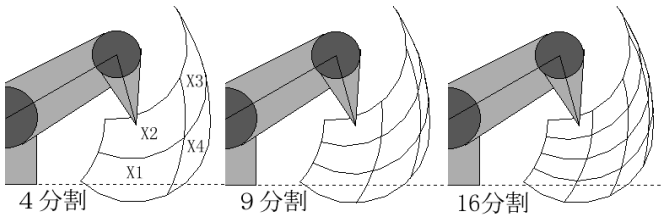

図 4: 連続値の角度センサ入力空間のグリッド分割による観測の離散化

を粗くすると POMDP になる．図 4のように，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．

3　確率的傾斜法による強化学習

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を提案した．従来研究では，何らかの状態推定器と Q-learning を組合わせるアプローチがほとんどだった。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る。しかし、状態推定を完全に保証することは一般に困難である上，状態推定器はしばしば非常に大きな状態空間を生成することがあり、何らかの汎化を行った等の理由により環境が MDP に変換できないため Q-learning では学習できない場合がある。よって本論文で考えるような最適化の必要性がある．確率的傾斜法は POMDP の環境における基本的な強化学習アルゴリズムである．以下，確率的傾斜法について説明する．

3.1　確率的傾斜法のエージェントの一般表現

エージェントの学習目標は，報酬獲得の平均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 5)．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．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．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．$\pi(a,W,X)$ 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．

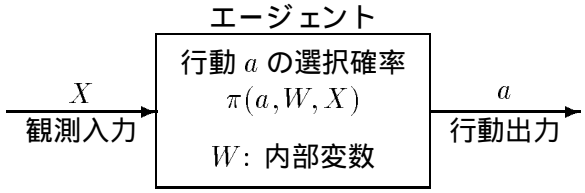

図 5: エージェントの確率的政策．内部変数 Wを調節することにより政策πを改善していく．

3.2　アルゴリズムの一般形

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 $(0 \le \gamma < 1)$ 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 6: 確率的傾斜法による強化学習法の一般形

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 6に示す．図 6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，どんな行動をとったのかについての情報である．$D_i(t)$ 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．

4 実験

4.1 実験設定

図 2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 ± 6 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 ± 6 度だけ回転する．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．

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 4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 7)．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．

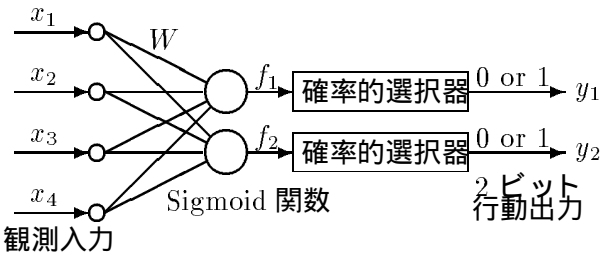

図 7: エージェントの内部構造 (離散入力)

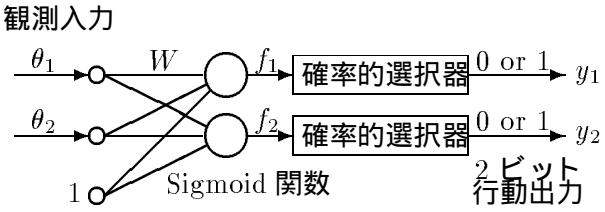

図 8: エージェントの内部構造 (連続値入力)

(2) 連続値の観測 (角度) を入力する場合 ： 各関節の角度を $0 \leq \theta_i \leq 1$ に正規化して入力する (図 8)．

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ユニットを結ぶ内部変数を 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．

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

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．

表 1: 行動選択確率と適正度

行動 a_j $= (y_1, y_2)$	行動選択確率 $\pi(a_j, W, X)$	適正度 $e(t) =$ $\frac{\partial}{\partial w} \ln \pi(a_j, W, X)$
$a_1 = (0,0)$	$(1-f_1)(1-f_2)$	$\frac{-1}{1-f_1}\frac{\partial f_1}{\partial w} + \frac{-1}{1-f_2}\frac{\partial f_2}{\partial w}$
$a_2 = (0,1)$	$(1-f_1)\, f_2$	$\frac{-1}{1-f_1}\frac{\partial f_1}{\partial w} + \frac{1}{f_2}\frac{\partial f_2}{\partial w}$
$a_3 = (1,0)$	$f_1\,(1-f_2)$	$\frac{1}{f_1}\frac{\partial f_1}{\partial w} + \frac{-1}{1-f_2}\frac{\partial f_2}{\partial w}$
$a_4 = (1,1)$	$f_1\, f_2$	$\frac{1}{f_1}\frac{\partial f_1}{\partial w} + \frac{1}{f_2}\frac{\partial f_2}{\partial w}$

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．

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

よって図 7,8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ユニットのローカルな情報だけを用いて計算可能となり，完全に並列分散処理可能である．式 (2) を図 6の一般形アルゴリズムにあてはめると，図 9のようになる．

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．

$$
\begin{aligned}
e_{ij}(t) &= x_i(y_j - f_j)\ , \\
D_{ij}(t) &= e_{ij}(t) + \gamma D_{ij}(t-1)\ ,
\end{aligned}
$$

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 9: エージェントに実装した確率的傾斜法

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，重み初期値は± 0.05 の範囲内でランダムに初期化する．

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 10は 100 試行の結果の平均を示す．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 11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 11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．

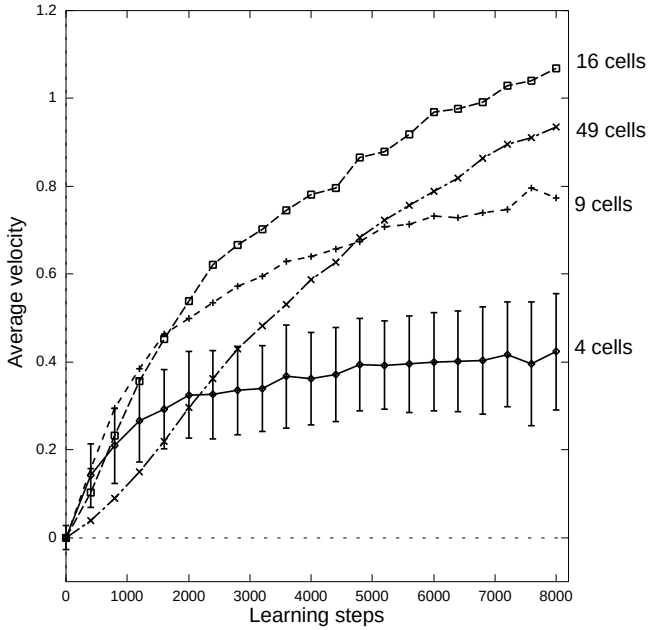

図 10: センサ (角度) 空間をグリッドで離散化入力した場合の学習結果．Cells は状態空間の分割個数，横軸は学習ステップ，縦軸は平均速度を表す．

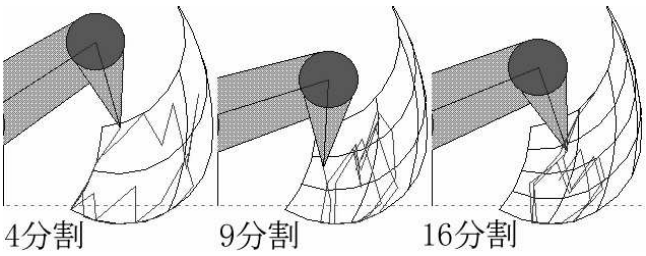

図 11: センサ (角度) 空間をグリッドで離散化入力した場合の学習で得たアーム先端の軌跡の一例．

観測入力が離散な場合，Q-learning[Watkins et.al 92] や Jaakkola の手法 [Jaakkola 94] と比較が可能である．図 12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ベストのチューニングを行い、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 の方法では割引率$\gamma = 0.95$ 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。

状態分割が荒く隠れ状態が存在する場合、図 12,13 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 15,16)。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。提案手法は分割数の変化に対して最もロバストであることがわかる。

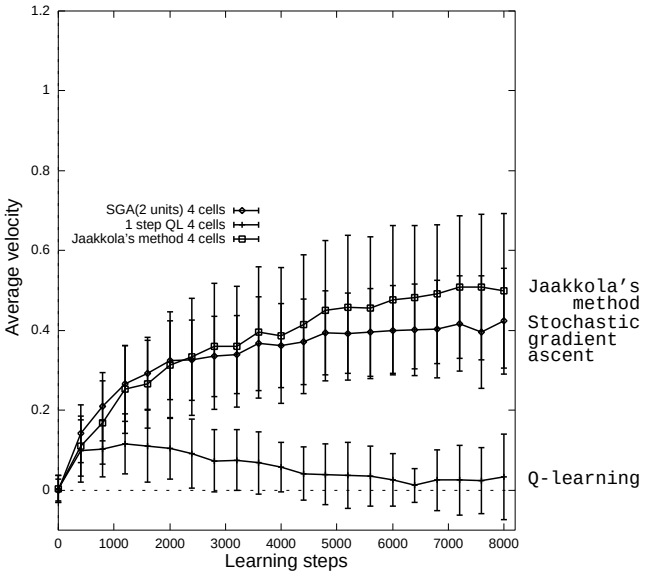

図 12: 角度センサの入力空間を 4 分割して離散化した場合の比較 . 横軸は学習ステップ , 縦軸は平均速度を表す。確率的傾斜法は SGA と記した。

(2) 連続値の観測 (角度) を入力した場合 : 図 17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 . 図 18 は学習で得た行動の例を示す . アームを上方へと無駄に動かしているが , 重み変数 6 コ , シグモイド関数 2 コという貧弱な計算資源では , これが限界だったためと考えられる . 限られた計算資源の範囲内で状態空間を適切に汎化し , それなりの政策を得た .

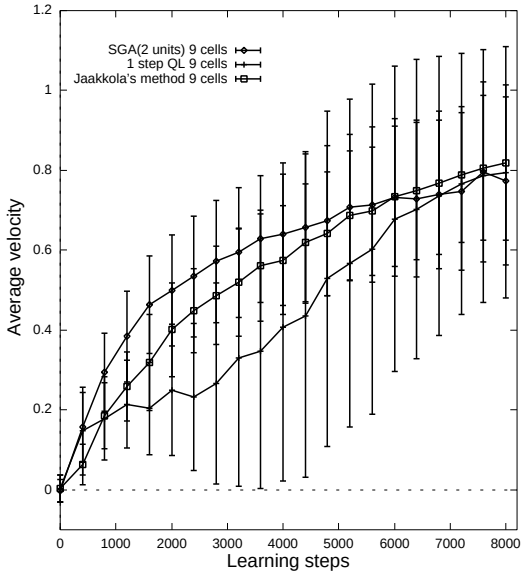

図 13: 角度センサの入力空間を 9 分割した場合の比較。

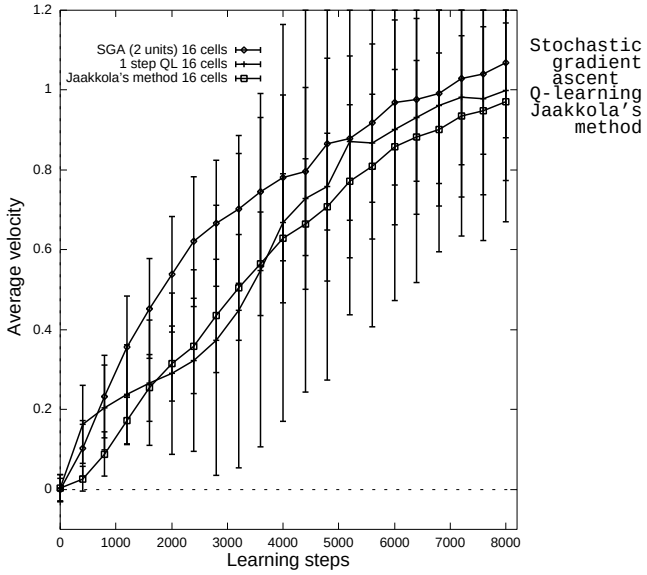

図 14: 角度センサの入力空間を 16 分割した場合の比較。

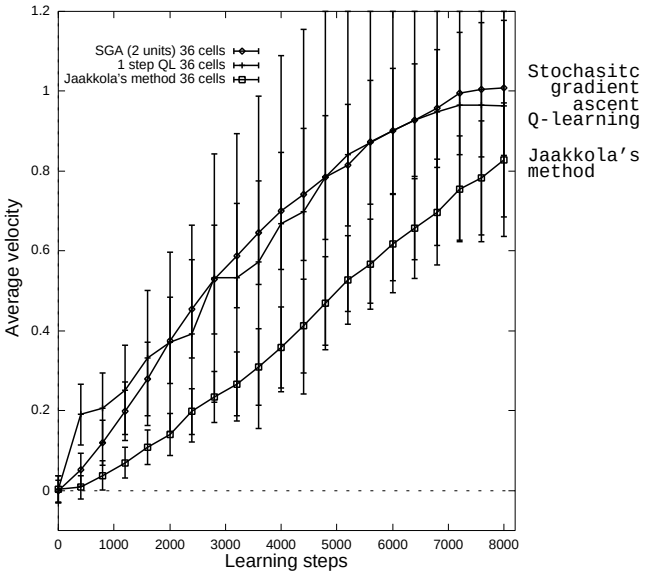

図 15: 角度センサの入力空間を 36 分割した場合の比較。

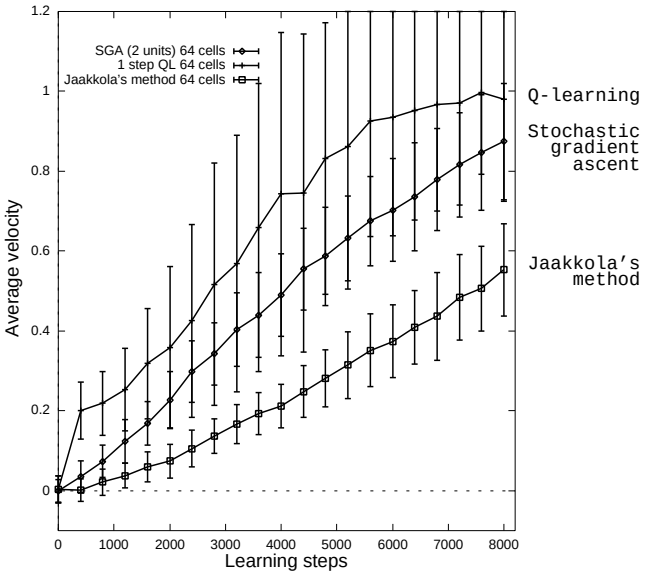

図 16: 角度センサの入力空間を 64 分割した場合の比較。Jaakkola の方法は状態空間の増加に対して敏感であることがわかる。

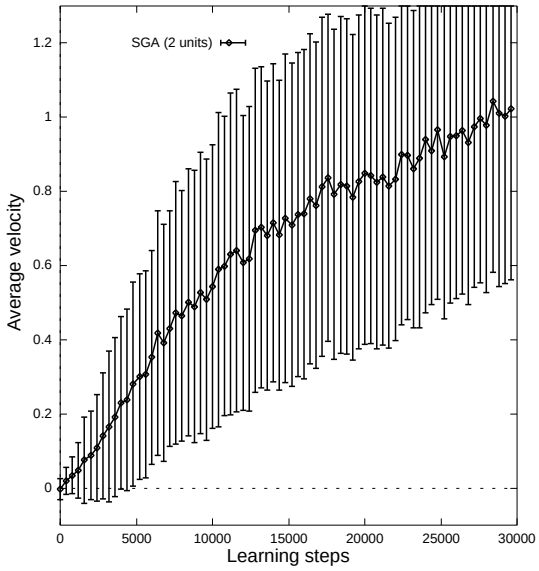

図 17: 連続値の観測 (角度) を入力した場合の学習結果．横軸は学習ステップ，縦軸は平均速度を表す．

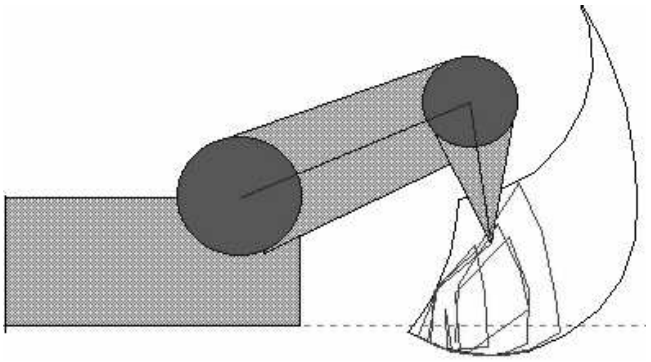

図 18: 連続値の観測 (角度) を直接入力した場合の学習で得たアーム先端の軌跡の一例．

5 考察

頑健性 :本手法は，状態空間の分割がたった 4 分割でマルコフ性を満たさない困難な場合においても，全くお手上げとはならずに半分程度の性能を得た．他手法と比較しても本手法は分割数の変化に対して最もロバストである．離散分割の時と全て同じ学習パラメータで，関節の角度を直接入力した場合でも，連続な状態空間を適切に汎化して学習した．よって本手法は，どのような構造のエージェントに適用してもその関数近似能力の範囲で性能を最大化するような頑健性を有していると考えられる．

探査戦略 :本手法は，訪れる頻度の高い所ほど優先的に学習が進む傾向を持っている．よって頻度の低い所では相対的にランダムに行動する傾向がある．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制約 :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で微分できること程度の条件さえ満たせばよいので，極めて多様な構造をとることができる．実験で用いた構造のエージェント (行動 4 コに対して重み変数 6 コ，シグモイド関数 2 コ相当の計算資源) へ Q-learning や Jaakkola の手法を適用することは不可能である．

確率的政策 :本手法は，4 分割の場合の領域 $X1$ のように，確率的な行動が必要な領域では，それを維持し続けることができる．Q-learning では学習が進行すると最大の Q 値を持つ行動だけを優先的に選択する探査戦略をとるのが一般的であり，上記のような振舞いは困難である．確率的政策は，マルチエージェント間のゲームなどへの適用が期待できる．

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．次にエージェントへ確率的傾斜法を実装してシミュレーションを行い，本手法が，どのような構造のエージェントに適用してもその関数近似能力の範囲で性能を最大化する頑健性を有していることを示した．

本論文では単純に観測入力から行動出力への確率分布を最適化するためのメモリーレスな強化学習法についてのみ議論したが，隠れ状態問題では何らかの状態推定手法と，メモリーレスな強化学習の 1 つである Q-learning を組み合わせるアプローチが一般的である．今後はこれらの状態推定手法との組み合わせによる POMDP へのアプローチを考えている．

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.

[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).

[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).

[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).