博士（工学）論文

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

Policy Improvement by Stochastic Gradient Ascent:

A New Approach to Reinforcement Learning in POMDPs

指導教官　　小林重信　教授

学籍番号　　９４Ｄ３１０３７

氏　　名　　木村　元

目 次

第 1 章　はじめに **3**

1.1　研究の背景 3

1.2　研究の目的と方法 4

1.3　論文の構成 5

第 2 章　問題設定および接近法 **6**

2.1　部分観測マルコフ決定過程 (POMDPs) 6

2.2　従来の接近法と問題点 7

2.3　強化学習エージェントの学習目標 9

2.3.1　確率的政策 9

2.3.2　最適性 10

2.4　POMDPs における数学的性質 11

2.4.1　直接報酬の期待値と政策 11

2.4.2　期間 N の割引報酬の期待値と政策 13

第 3 章　政策の逐次改善アルゴリズム:　確率的傾斜法の提案と解析 **14**

3.1　アルゴリズムの提案 14

3.2　アルゴリズムの解析 16

3.3　アルゴリズムの動作例 18

3.3.1　実験設定 18

3.3.2　実験の結果と解析 21

3.4　アルゴリズムの特徴のまとめ 24

第 4 章　ロボットアームの制御問題への適用 **25**
4.1 ロボットアームの制御問題 25
4.2 実験設定 29
4.3 エージェントの実装 29
4.3.1 角度センサ入力空間をグリッドで分割/離散化する場合 29
4.3.2 連続値の観測 (角度) を入力する場合 30
4.3.3 確率的傾斜法の実装 30
4.4 実験結果 32
4.4.1 センサ (角度) 空間をグリッドで分割した場合 32
4.4.2 連続値の観測 (角度) を入力した場合 39
4.5 考察 41

第 5 章　結　論 **42**
5.1 研究成果のまとめ 42
5.2 今後の展望 43

付録 A　定理 2 の証明 **44**

謝辞 **52**

公表論文 **53**

参考文献 **54**

第 1 章

はじめに

1.1　研究の背景

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 1.1)。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。

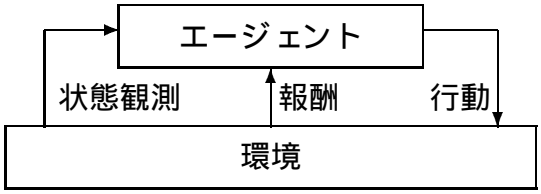

図 1.1: 強化学習の枠組

例えば、道路の真ん中で「ロボットの足が壊れた」という状況を考える。歩けないからといって、じっとしていたら車にはねられてしまうので、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。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た、報酬と呼ばれる一種の評価値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。一般に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。

1.2 研究の目的と方法

従来の多くの強化学習研究はマルコフ決定過程 (MDPs) を対象としてきたが、現実の問題の多くはこのような仮定を満たさない。実世界における強化学習では、隠れ状態 (hidden state) と関数近似の 2 つの問題の扱いが求められる。連続で大きな状態空間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 [Lin 93][Tan 91][Tesauro 92]。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 [Singh et.al 94]。例えば連続な状態空間において離散化が荒すぎる場合などに生じる [Moore et.al 95]。また、[Whitehead et al. 95] の不完全知覚 (Perceptual aliasing) もこれと同じと考えられる。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。本論文では隠れ状態問題を部分観測マルコフ決定過程 (partially observable Markov decision processes: POMDPs) として一般化する [Singh et.al 94]。以上より、実問題におけるエージェントは、POMDP の環境において関数近似を用いて学習を行う必要があると考えられる。

本論文では、POMDP の環境において、関数近似されたメモリレスな政策、すなわち観測から行動への写像を形成する強化学習アルゴリズムを提案する。提案手法は各時間ステップでの状態や報酬の期待値等を明示的に推定するような計算コストのかかる処理が不

用であり、実時間処理に向いた方法である。

1.3 論文の構成

以下第 2章では問題設定の詳細および従来の接近法とその問題点について述べる。第 3章では、2章で示した性質を利用した強化学習法を提案する。また、提案手法の持つ性質を理論的に解析し、実験により確認する。第 4章では提案したアルゴリズムをロボットアームの制御問題へ適用し、他手法との比較などによりその性質につて考察する。最後に第 5章では本研究の成果についてまとめる。

第 2 章

問題設定および接近法

2.1 部分観測マルコフ決定過程 (POMDPs)

強化学習における部分観測マルコフ決定問題は以下のように定義される。エージェントが接する環境には、基礎となるマルコフ決定過程 (MDP) が存在している。この基礎となる MDP の状態集合を $\mathcal{S} = \{s_1, s_2, \cdots, s_n\}$, エージェントが選択可能な行動の集合を $\mathcal{A} = \{a_1, a_2, \cdots, a_l\}$ と表す。状態 s において行動 a を実行したとき、状態 s'へ遷移する確率を $P^a(s, s')$ と表す。状態 s において行動 a を実行したときエージェントが受け取る報酬の期待値を $R^a(s)$ と表す。エージェントは MDP の状態 s を直接知ることはできないが、状態についての情報を含んだ "観測" を受けとる。観測の集合を $\mathcal{X} = \{X_1, X_2, \cdots, X_m\}$ と表す。MDP における状態が s のとき、エージェントは $P(X|s)$ の確率で観測 Xを受けとる。$P(X|s)$ は現在の状態 s だけに依存するものとする。

ある時間ステップ t においてエージェントは環境中の状態 s_tにある。このときエージェントは $P(X_t|s_t)$ の確率で環境から観測 X_tを受けとり、行動 a_tを実行する。その結果、エージェントは期待値 $R^{a_t}(s_t)$ で環境から報酬 r_tを受けとり、環境は確率 $P^{a_t}(s_t, s_{t+1})$ に従って状態 s_{t+1}へ遷移し、次の時間ステップ $t+1$ へと進む。

エージェントは基礎となる MDP や $P(X|s)$, $R^{a_t}(s_t)$ に関する知識をあらかじめ持っていないものとする。

2.2 従来の接近法と問題点

従来の強化学習における非マルコフな環境へのアプローチのほとんどは、環境における真の状態を推定する手法と、推定された状態から行動を決定するための学習との組み合わせを基本としている。状態の推定では、以下のように様々な手法が提案されている。[Lin et.al 92] や [Whitehead et al. 95] の Stored-state methods や、[Chrisman 92] や [McCallum 95b] の instance-based methods は、基本的にエージェントの観測-行動の時系列を現在の状態とする。環境のモデルを必要とする強化学習 (model-based RL) では、自分がどの状態にいるのかを表す確率分布 (belief state) を状態表現として用いる方法が提案されている [Littman et.al 95]。この状態推定を行うには、一般にいろいろな仮定が必要になる。残念ながら実際問題ではエージェントの計算資源の制約などにより、この仮定を満たさない場合が多い。そのため、正確な状態推定を保証することが困難な場合がある。よって、POMDP において memory-less な確率的政策を求めるような強化学習法が必要である。また、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るので、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。

推定された状態から行動を決定するための学習法としては、ほとんどの場合ダイナミックプログラミングを基礎とする TD 法 [Sutton 88] や Q-learning [Watkins et.al 92] が用いられている。これらの手法は Q-net[Lin et.al 92] のように関数近似と組み合わせることができる。MDP の環境においては、DP に基づく強化学習に関数近似を用いた場合の理論的な解析が示されているが [Baird 95b] [Boyan et.al 94] [Gordon 95] [Heger 96] [Singh et.al 94]、POMDP の環境では理論的解析は示されていない。Q(λ) learning [Peng et.al 94] や TD(λ) [Sutton 95] がある種の非マルコフ問題において有効であることが実験的に示されている。これら DP に基づく学習は MDP において決定的な政策を得るための手法なので、状態遷移がマルコフ性を満たさない場合には一般に問題が生じる [Singh et.al 94]。そのため、状態推定が正確になるまでの学習途中や、正確な状態推定が保証できない場合などは、DP に基づく学習を用いた政策改善には限界がある。

離散 POMDPs における効率のよい強化学習法として、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 et.al 94]。しかしながら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。

[Williams 92] は、報酬に遅れのある問題において報酬の期待値の勾配の方向へパラメータを更新していく episodic REINFORCE アルゴリズムを提案した。このアルゴリズムの理論的解析結果は POMDPs の環境においても成り立ち、さらにこのアルゴリズムは、関数近似された memory-less な政策を改善していくものである。よって Williams の方法が従来の強化学習アルゴリズムの中で最も有望だと考えられるが、エルゴート性を有する POMDPs におけるエピソードの定義と政策の更新方法には問題があった。

本論文で提案するアルゴリズムは、Williams の方法を拡張したものである。本手法では、Williams の REINFORCE における適正度 (eligibility) を割引きながら足し合わせた適正度の履歴 (eligibility trace) を用いて政策を改善する。これにより、本手法の政策更新は、エピソード毎のバッチ的な処理とは異なり、毎ステップにおいて逐次的に行われる。

2.3 強化学習エージェントの学習目標

2.3.1 確率的政策

エージェントの学習目標は、何らかの報酬関数を最大化する政策つまり観測から行動への写像を形成することである。政策は、それぞれの観測において行動を選択する確率分布を割り当てる。この政策πのもとで観測Xにおいて行動aを選択する確率$Pr(a|\pi, X)$は、エージェントの内部変数ベクトルWを用いて(2.1)式の確率密度関数で表す。

$$Pr(a|\pi, X) \;=\; \pi(a, W, X)\ . \tag{2.1}$$

行動選択確率を計算する機構が、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するものであり、重み付きのルールベースシステムならば、内部変数Wはルールの重みに相当するものである。エージェントは内部変数Wを調節することで政策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。本研究では、このように様々な構造を持つと考えられる任意のエージェントの構造および制約条件を単純に$\pi(a, W, X)$の関数形で表している。よって、そのすべてのエージェントに対して理論的な基礎を与えることができる。これが本研究によるアプローチの大きな特徴である。

内部変数Wを固定して政策πを定常とすると、POMDPの系全体の確率法則が全て定まる。このときの定常政策πのもとでの状態sからs'への遷移確率を$P^{\pi}(s, s')$とし、(2.2)式に定義する。

$$P^{\pi}(s, s') = \sum_{a \in \mathcal{A}} \sum_{X \in \mathcal{X}} P(X|s)\, \pi(a, W, X)\, P^{a}(s, s')\ . \tag{2.2}$$

本論文では、全ての政策πのもとで基礎となるマルコフ過程はエルゴート性を有するものとする。

2.3.2 最適性

エージェントが受けとる報酬の平均の期待値を平均報酬Λと表し、以下の式で定義する。

$$\Lambda = \lim_{N\to\infty} E\Big\{\frac{\sum_{t=0}^{N} r_t}{N+1}\Big\}\ , \tag{2.3}$$

ただし $E\{\cdot\}$ は期待値を表す。定常政策πのエージェントの平均報酬をΛ^{π}と表し、これを最大化する政策を最適政策と定義する。基礎となるマルコフ過程は全ての政策πのもとでエルゴート性を仮定しているので、平均報酬Λ^{π}は初期状態には依存しない。POMDPs の環境においてメモリーレスな方法で学習する場合、平均報酬による最適性の定義が最も適切であることが示されている [Singh et.al 94]。エージェントの学習目標は、平均報酬Λ^{π}を最大化するような定常政策を形成することである。

2.4 POMDPs における数学的性質

本章では POMDP の性質についてまとめ、新しい強化学習法を導く上で有用な知見を与える。

2.4.1 直接報酬の期待値と政策

状態 s におけるエージェントが定常政策πのもとで行動を選択した結果、即座に受け取る報酬の期待値を $R^{\pi}(s)$ と表し、(2.4) 式に定義する。

$$R^{\pi}(s) = \sum_{a \in \mathcal{A}} \sum_{X \in \mathcal{X}} P(X|s)\, \pi(a, W, X)\, R^{a}(s)\ . \tag{2.4}$$

$\forall X \in \mathcal{X}$, $\forall a \in \mathcal{A}$ 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。W の任意の i 番目の要素を w_i と表す。それぞれの時間ステップ t において、エージェントは (2.5) 式で定義される $e_i(t)$ を容易に計算できる。

$$e_i(t) \ = \ \frac{\partial}{\partial w_i} \ln\Big(\pi(a_t, W, X_t)\Big)\ . \tag{2.5}$$

(2.5) 式の $e_i(t)$ は適正度 $(\text{characteristic eligibility})$ と呼ばれ [Williams 92]、実行した行動についての情報論的な意味での価値を表す。例えば選択確率の小さな行動をとった場合、情報としての価値は大きいので適正度 $e_i(t)$ は大きな値となる。エージェントは行動選択確率πの計算途中で $e_i(t)$ を容易に得ることができる。このとき、以下の定理が成り立つ。

定理 1 *POMDP* の環境において、

$$E\{(r_t - b)\, e_i(t)\} \ = \ \frac{\partial}{\partial w_i} R^{\pi}(s_t)\ , \tag{2.6}$$

ただし b は *reinforcement baseline* と呼ばれる定数、r_tは時間ステップ t においてエージェントが受けとる報酬である。

定理 1の証明: (2.6) 式の左辺は確率論の公理より

$$E\{(r_t - b)\, e_i(t)\} \;=\; E\{r_t\, e_i(t)\} - E\{b\, e_i(t)\}\ . \tag{2.7}$$

ここで右辺の第 2 項に注目する。(2.7). (2.5) 式より、

$$\begin{aligned}
E\{b\, e_i(t)\} &= E\Big\{b\,\frac{\partial}{\partial w_i}\ln(\pi(a_t, W, X_t))\Big\} && (2.8)\\
&= \sum_{X\in\mathcal{X}}\sum_{a\in\mathcal{A}} b\, P(X|s_t)\,\pi(a, W, X)\frac{\partial}{\partial w_i}\ln(\pi(a, W, X)) && (2.9)\\
&= b\sum_{X\in\mathcal{X}}\sum_{a\in\mathcal{A}} P(X|s_t)\frac{\partial}{\partial w_i}\pi(a, W, X) && (2.10)\\
&= b\,\frac{\partial}{\partial w_i}\Big\{\sum_{X\in\mathcal{X}}\sum_{a\in\mathcal{A}} P(X|s_t)\,\pi(a, W, X)\Big\} && (2.11)\\
&= b\,\frac{\partial}{\partial w_i}1 \;=\; 0\ . && (2.12)
\end{aligned}$$

(2.12) 式までの導出について補足する。時間ステップ t において行動 a を選択する確率は $P(X|s_t)\,\pi(a, W, X)$ である。ただし s_tはこのときの基礎となる MDP の状態を示す。よって (2.8) 式の示すの期待値は (2.9) 式で表される。 (2.10) 式の変形には対数関数の微分公式を用いた。$P(X|s_t)$ は政策πに対して独立なので (2.11) 式へ変形できる。同様にして

$$\begin{aligned}
E\{(r_t - b)\, e_i(t)\} &= E\{r_t\, e_i(t)\} \;-\; 0\\
&= E\Big\{r_t\,\frac{\partial}{\partial w_i}\ln(\pi(a_t, W, X_t))\Big\}\\
&= \sum_{X\in\mathcal{X}}\sum_{a\in\mathcal{A}} R^a(s_t)\, P(X|s_t)\,\pi(a, W, X)\frac{\partial}{\partial w_i}\ln(\pi(a, W, X))\\
&= \sum_{X\in\mathcal{X}}\sum_{a\in\mathcal{A}} R^a(s_t)\, P(X|s_t)\frac{\partial}{\partial w_i}\pi(a, W, X)\\
&= \frac{\partial}{\partial w_i}\Big\{\sum_{X\in\mathcal{X}}\sum_{a\in\mathcal{A}} P(X|s_t)\,\pi(a, W, X)\, R^a(s_t)\Big\} && (2.13)\\
&= \frac{\partial}{\partial w_i} R^\pi(s_t)\ . && (2.14)
\end{aligned}$$

(2.14) 式までの導出について補足する。$P(X|s_t)$, $R^a(s_t)$ は政策πに対して独立なので (2.13) 式へ変形できる。(2.13) 式は (2.4) 式を用いると (2.14) 式で表される。 ■

定理 1は、POMDP の環境に置かれた強化学習エージェントの政策改善方法に関して重要な手がかりを与える。エージェントは毎ステップにおいて容易に計算できる (2.5) 式と

報酬 r_tだけから、報酬の期待値 $R^\pi(s_t)$ を最も大きくする方向へ政策πを確率的に改善することが可能である。ここで注目すべき特徴は、現在の状態 s_tや報酬の期待値 $R^\pi(s_t)$ を明示的に推定する必要がないということである。

2.4.2 期間 Nの割引報酬の期待値と政策

将来における報酬が現在の評価に与える影響を決める定数を割引率γで表す。ただし $0 \leq \gamma < 1$ である。定常政策πをとり続けた場合における状態 s_0の期間 Nの割引報酬の期待値 $V_N^\pi(s_0)$ は (2.2),(2.4) 式より (2.15) 式で表される。

$$
\begin{aligned}
& V_N^\pi(s_0) \\
= \;& R^\pi(s_0) + \gamma \Big\{ \sum_{s_1 \in \mathcal{S}} P^\pi(s_0, s_1)\, R^\pi(s_1) \Big\} + \gamma^2 \Big\{ \sum_{s_1 \in \mathcal{S}} \sum_{s_2 \in \mathcal{S}} P^\pi(s_0, s_1)\, P^\pi(s_1, s_2)\, R^\pi(s_2) \Big\} + \cdots \\
& + \gamma^t \Big\{ \sum_{s_1 \in \mathcal{S}} \sum_{s_2 \in \mathcal{S}} \cdots \sum_{s_{t-1} \in \mathcal{S}} \sum_{s_t \in \mathcal{S}} P^\pi(s_0, s_1)\, P^\pi(s_1, s_2) \cdots P^\pi(s_{t-1}, s_t)\, R^\pi(s_t) \Big\} + \cdots \\
& + \gamma^N \Big\{ \sum_{s_1 \in \mathcal{S}} \sum_{s_2 \in \mathcal{S}} \cdots \sum_{s_{N-1} \in \mathcal{S}} \sum_{s_N \in \mathcal{S}} P^\pi(s_0, s_1) P^\pi(s_1, s_2) \cdots P^\pi(s_{N-1}, s_N)\, R^\pi(s_N) \Big\} \; . \quad (2.15)
\end{aligned}
$$

なお、割引報酬と平均報酬の間には、以下の関係が成り立つ。(2.15) 式において、割引率 $0 \leq \gamma < 1$ で，期間 $N \to \infty$ の極限の場合を考える。このときの政策πにおける状態 s の極限分布 (占有確率) を $U^\pi(s)$ と表す。すると平均報酬Λ^πは (2.16) 式のように表される [Singh et.al 94]。

$$
\frac{\Lambda^\pi}{1-\gamma} \;=\; \sum_{s \in \mathcal{S}} U^\pi(s) V_\infty^\pi(s) \quad \text{, where } 0 \leq \gamma < 1 \; . \qquad (2.16)
$$

第 3 章

政策の逐次改善アルゴリズム: 確率的傾斜法の提案と解析

3.1 アルゴリズムの提案

平均報酬を最大化するように政策を改善するには、episodic REINFORCE アルゴリズムによって、十分大きな期間 Nおきに報酬の平均を計算して政策を更新するバッチ処理的な方法が考えられる。しかし強化学習では上記のようなバッチ処理的な方法よりも、特定の時間ステップに処理が集中せず毎ステップ同じ処理を繰り返して逐次的に政策を更新していく方が好ましい。そこで、POMDP の環境において逐次的に政策を改善する強化学習法を提案する (図 3.1)[Kimura et.al 95]。

図 3.1 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 $e_i(t)$ は定理 1と同じで適正度 (eligibility) と呼ばれ、実行した行動についての情報論的な意味での価値を表す量である。例えば選択確率の小さな行動をとった場合、情報としての価値は大きいので適正度 $e_i(t)$ は大きな値となる。エージェントは行動選択確率πの計算途中で $e_i(t)$ を容易に得ることができる。$D_i(t)$ は適正度の履歴 (eligibility trace) であり、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。割引率 $\gamma = 1$ に設定すると $D_i(t)$ が発散してしまうため $0 \leq \gamma < 1$ でなければならない。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し合い、報酬獲得に関係する行動だけが強化されていく。行動の履歴を強化するので、

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。

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

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

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

   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、手順 1 へ戻る。

図 3.1: 政策の逐次改善アルゴリズム :確率的傾斜法

報酬の獲得に遅れのある行動も強化される。過去の経験を割り引き率で減衰させる理由は以下による。学習の進行によって、時間的に古い過去の政策ほど現在の政策との違いが大きくなっていく。そのため、現在の政策改善に対して過去の経験を利用するにあたり、現在の政策と過去の経験時の政策との違いを考慮する必要が生じるからである。

3.2 アルゴリズムの解析

本研究では提案したアルゴリズムについて以下に示す解析結果を得た。

定理 2 全ての時間ステップ $t \geq 0$ において報酬の絶対値 $|r_t|$ が上界を持ち、また全ての $a \in \mathcal{A},\ X \in \mathcal{X},\ W$ において $|\frac{\partial}{\partial w_i} \ln \pi(a, W, X)|$ が上界を持つものとする。$POMDP$ の環境において、提案したアルゴリズムが定常な政策πを保ったままのとき、, ΔWは以下の式を満たす。

$$\lim_{N\to\infty} \frac{1}{(1-\gamma)N} \sum_{t=0}^{N-1} \Delta w_i(t) \ = \sum_{s\in S} U^{\pi}(s) \frac{\partial}{\partial w_i} V_{\infty}^{\pi}(s) \ , \tag{3.1}$$

ただし $U^{\pi}(s)$ は政策πにおける状態 s の極限分布 *(*占有確率*)* を表し、割引報酬の割引率はアルゴリズム中で用いる割引率γと等しいものとする。

この定理では、提案手法における $\Delta w_i(t)$ の平均値が状態の極限分布 (占有確率) に重み付けされた割引報酬の期待値の傾斜に等しいことを示している。これより、提案手法が割引報酬に関する確率的傾斜法になっていることが分かる。証明は付録 A に示す。ここで注目すべき特徴は、図 3.1 のアルゴリズムでは状態の明示的な推定を行っていないのにもかかわらず、状態に関する割引報酬を改善している点である。

この定理を利用することにより、本アルゴリズムがどのような解に収束するのかを予測することができる。POMDP の環境において，提案したアルゴリズムが定常な政策πに収束したとき、ΔWは (3.1) 式を満たす。

またこの定理では、今までとった行動の履歴である $D_i(t)$ を割引率γで減衰することが割引率γの割引報酬に深く関係していることを示している。離散 MDP の環境において、ルックアップテーブルを用いたエージェントに本アルゴリズムを適用すると、割引率 γの割引報酬を極大化する政策を得る。学習途中のような非定常な場合においては、時間的に古い過去の政策ほど現在の政策との違いが大きくなっていくので、適正度の履歴 $D_i(t)$ を割引率γで割引くことが平均報酬の改善に効果的であると期待できる。しかし政策が定常な場合においては、平均報酬Λ^{π}に関する (2.16) 式より、状態 s の極限分布 (占有確率) が

政策πやw_iに依存する場合には一般に以下のようになる。

$$\frac{\partial}{\partial w_i}\frac{\Lambda^\pi}{1-\gamma} \;=\; \frac{\partial}{\partial w_i}\sum_{s\in\mathcal{S}} U^\pi(s)V_\infty^\pi(s) \quad \neq \quad \sum_{s\in S} U^\pi(s)\frac{\partial}{\partial w_i}V_\infty^\pi(s)$$

よって割引率γを一定の値に固定したまま学習が進んでほぼ定常な政策に収束したとき、これは残念ながら平均報酬を極大化する定常政策にはならない。しかしながら、割引率γを1に近付けることにより、本アルゴリズムの得る政策を平均報酬の極大化政策に近付けることができる。

この定理ではまた、本アルゴリズムが存在確率の高い状態ほど優先的に割引報酬を改善していくことを示している。これは、本アルゴリズムの単純な処理の中に合理的な探査戦略を行うメカニズムが含まれていることを意味する。例えば、よく訪れる状態についてはほぼ最適な決定的行動選択をするように収束していても、めったに訪れることのない状態に直面した場合には確率的に行動選択すると考えられる。

3.3 アルゴリズムの動作例

提案した強化学習アルゴリズムが定理 2 の (3.1) 式に示した政策に収束することを確認する。また、割引率 γ を 1 に近付けると平均報酬を極大化する定常政策に近付くことを確認する。

3.3.1 実験設定

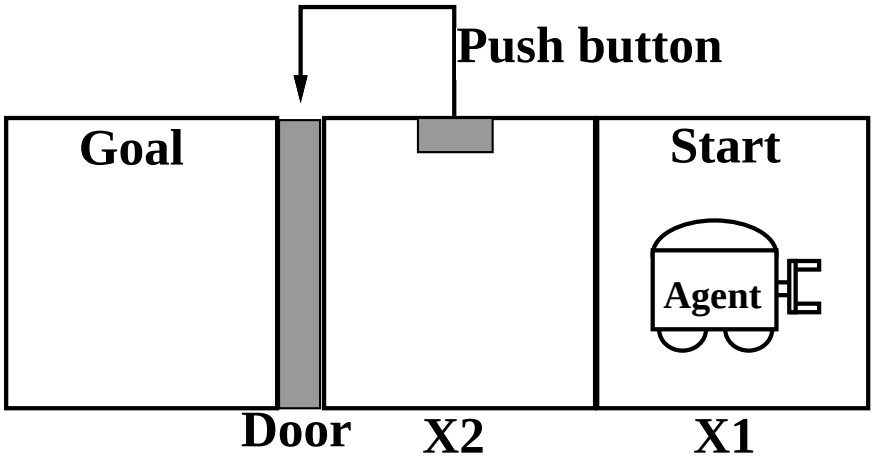

図 3.2: 実験に用いた環境。ボタンを押すと確率 $1-e$ でドアが開く。

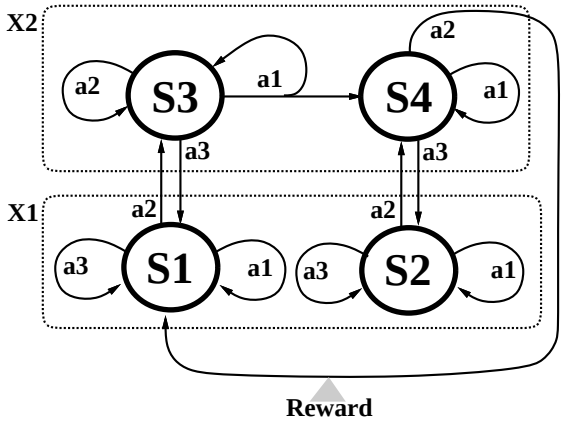

図 3.3: 実験に用いた環境の状態遷移図。S_3 で a_1 を実行すると確率 $1-e$ で S_4 へ遷移する。

解析を簡単にするため、図 3.2 に示す環境を考える。エージェントは自分の位置として

X_1またはX_2を観測する。エージェントは、左に動く、右に動く、ボタンを押すの３種類の行動を選択できる。初期状態でエージェントはスタート位置X_1に配置され、ドアは閉じている。X_2の位置でエージェントがボタンを押すと、確率$1-e$でドアが開く。一度ドアが開けば、エージェントがドアを通らない限りドアは開いた状態を保持する。エージェントはドアの開閉状態を観測できない。ドアが開いている状態においてX_2の位置でエージェントが左に動けば、エージェントは報酬を受けとってスタート位置X_1へ戻り、ドアは閉まる。図3.2の環境は、4つの状態$S_1,\ S_2,\ S_3,\ S_4$と2つの観測$X_1,\ X_2$および３種類の行動$a_1 = Push,\ a_2 = Left,\ a_3 = Right$からなるPOMDPでモデル化できる。図3.3は図3.2の環境をPOMDPで表した状態遷移図である。図中の円は状態、矢印は行動による状態遷移を表す。S_2でa_1を実行すると、確率$1-e$でS_4へ遷移する。S_4でa_2を実行すると。エージェントは報酬1を受けとってS_1へ遷移する。

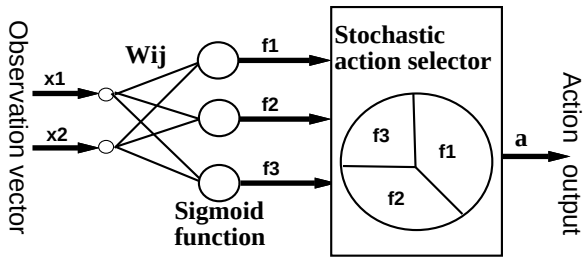

図 3.4: 実験に用いたエージェントの内部構造

実験に用いたエージェントの内部構造を図3.4に示す。観測Xは入力ユニット$x_1,\ x_2$に観測ベクトルとして入力される。エージェントの観測がX_1のとき、つまり図3.3での状態がS_1またはS_2のとき$(x_1, x_2) = (1, 0)$の値が入力される。観測がX_2のとき$(x_1, x_2) = (0, 1)$となる。ここでは任意のi番目の入力ユニットと任意のj番目のシグモイド関数ユニットを結ぶ内部変数をw_{ij}のように表す。これに伴い、対応するcharacteristic eligibilityを$e_{ij}(t)$, eligibility traceを$D_{ij}(t)$と表す。入力ユニットx_iと内部変数W_{ij}の加重和をとり、シグモイド関数ユニットからそれぞれの行動a_jに関する評価値$f_1,\ f_2,\ f_3$を以下のように計算する。

$$f_1 \quad = \quad \frac{1}{1 + \exp(-x_1 w_{11} - x_2 w_{21})}\ ,$$

$$
\begin{aligned}
f_2 &= \frac{1}{1+\exp(-x_1 w_{12} - x_2 w_{22})} \ , \\
f_3 &= \frac{1}{1+\exp(-x_1 w_{13} - x_2 w_{23})} \ .
\end{aligned}
$$

$f_1,\ f_2,\ f_3$ の割合に比例した確率で行動 $a_1,\ a_2$ または a_3 を選択する。よって政策 π は以下のように表される。

$$
\begin{aligned}
\pi(a_1, W, X) &= \frac{f_1}{f_1+f_2+f_3} \ , \\
\pi(a_2, W, X) &= \frac{f_2}{f_1+f_2+f_3} \ , \\
\pi(a_3, W, X) &= \frac{f_3}{f_1+f_2+f_3} \ . \qquad (3.2)
\end{aligned}
$$

行動 $a_t = a_k$ を選択した場合の characteristic eligibility は (2.5) と (3.2) 式より以下のようになる。

$$
\begin{aligned}
e_{ik}(t) &= \frac{f_1+f_2+f_3-f_k}{f_k(f_1+f_2+f_3)} \frac{\partial}{\partial w_{ik}} f_k \ , \\
e_{ij}(t) &= \frac{-1}{(f_1+f_2+f_3)} \frac{\partial}{\partial w_{ij}} f_j \ , \text{where } j \neq k. \qquad (3.3)
\end{aligned}
$$

(3.3), (3.3) 式を図 3.1 のアルゴリズムにあてはめることで $D_{ij}(t)$, $\Delta w_{ij}(t)$ を計算して政策を更新する。学習定数 $\alpha = 0.4$ に設定し、内部変数 w_{ij} は ± 0.05 の範囲内でランダムに初期化しておく。

3.3.2 実験の結果と解析

図 3.3 の平均報酬を最大化する最適政策はマルコフ解析を用いて以下のように計算できる。

$$
\begin{aligned}
\pi(a_1, W, X_1) &= 0\ , \\
\pi(a_2, W, X_1) &= 1\ , \\
\pi(a_3, W, X_1) &= 0\ , \\
\pi(a_1, W, X_2) &= \begin{cases} \frac{1-\sqrt{1-e}}{e} & , \text{where } e \neq 0, \\ \frac{1}{2} & , \text{where } e = 0, \end{cases} \\
\pi(a_2, W, X_2) &= \begin{cases} \frac{e-1+\sqrt{1-e}}{e} & , \text{where } e \neq 0, \\ \frac{1}{2} & , \text{where } e = 0, \end{cases} \\
\pi(a_3, W, X_2) &= 0\ .
\end{aligned}
$$

提案手法によって得る政策は、定理 2 を用いて以下のように予想される。学習によってほぼ定常な政策πに収束したとき、提案手法の性質を示す (3.1) 式より、

$$
\sum_{s \in S} U^{\pi}(s) \frac{\partial}{\partial w_i} V_{\infty}^{\pi}(s) = \left(\frac{\partial\, \pi(a, W, X)}{\partial w_i} \right) \sum_{s \in S} U^{\pi}(s) \frac{\partial}{\partial \pi(a, W, X)} V_{\infty}^{\pi}(s) \tag{3.4}
$$

ここで図 3.3 のマルコフ解析により、定常政策πにおいて常に以下の式が成り立つ。

$$
\begin{aligned}
\sum_{s \in S} U^{\pi}(s) \frac{\partial}{\partial \pi(a_1, W, X_1)} V_{\infty}^{\pi}(s) &< 0\ , \\
\sum_{s \in S} U^{\pi}(s) \frac{\partial}{\partial \pi(a_2, W, X_1)} V_{\infty}^{\pi}(s) &> 0\ , \\
\sum_{s \in S} U^{\pi}(s) \frac{\partial}{\partial \pi(a_3, W, X_1)} V_{\infty}^{\pi}(s) &< 0\ , \\
\sum_{s \in S} U^{\pi}(s) \frac{\partial}{\partial \pi(a_3, W, X_2)} V_{\infty}^{\pi}(s) &< 0\ .
\end{aligned}
$$

本アルゴリズムは上記の式の示す方向に政策を改善していくため、$\pi(a_1, W, X_1) \rightarrow 0$, $\pi(a_2, W, X_1) \rightarrow 1$, $\pi(a_3, W, X_1) \rightarrow 0$, $\pi(a_3, W, X_2) \rightarrow 0$ へ収束すると予想できる。また、

$$
\begin{aligned}
\sum_{s \in S} U^{\pi}(s) \frac{\partial}{\partial \pi(a_1, W, X_2)} V_{\infty}^{\pi}(s) &= 0\ , \\
\sum_{s \in S} U^{\pi}(s) \frac{\partial}{\partial \pi(a_2, W, X_2)} V_{\infty}^{\pi}(s) &= 0\ .
\end{aligned}
$$

を満たす$\pi(a_3, W, X_2)$ と $\pi(a_1, W, X_2)$ を求めると、

$$\begin{aligned} \pi(a_1, W, X_2) &= \begin{cases} \frac{1+\gamma-\sqrt{(1+\gamma)^2-4e\gamma^2}}{2e\gamma} & , \text{where } e \neq 0, \\ \frac{\gamma}{1+\gamma} & , \text{where } e = 0, \end{cases} \\ \pi(a_2, W, X_2) &= 1 - \pi(a_1, W, X_2) \ . \end{aligned}$$

となるので、政策$\pi(a_3, W, X_2)$ と $\pi(a_1, W, X_2)$ は上記の停留点へ収束すると予想される。

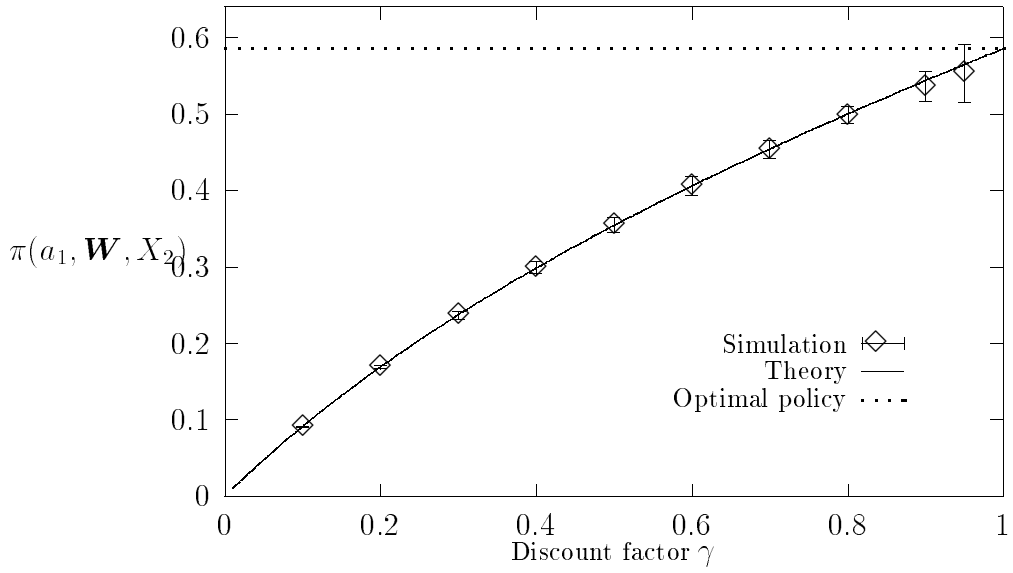

図 3.5: 提案手法が図 2 の環境において得る政策の実験結果と理論値。シミュレーションは 100 回行い、50000 ステップ学習後の平均と標準偏差をとった。割引率を 1 に近付けると最適政策に近付く。

ドアを開けるのに失敗する確率 $e = 0.5$ に固定し、割引率 γ を変化させた場合において得られる政策のシミュレーション結果を示す。図 3.5 は割引率 γ を変化させたときに得られる政策$\pi(a_1, W, X_2)$ の理論値とシミュレーション結果である。図中の点線は平均報酬を最大化する最適政策を表す。またシミュレーションでは、理論による予想通りに $\pi(a_1, W, X_1) \to 0$, $\pi(a_2, W, X_1) \to 1$, $\pi(a_3, W, X_1) \to 0$, $\pi(a_3, W, X_2) \to 0$ へ収束した。理論値とシミュレーション結果は一致している。これより、提案手法によって収束する政策は (3.1) 式により予測できることが確認された。また割引率 γを 1 に近付けると平均報酬

を極大化する定常政策に近付くことも確認された。よって学習途中において γ を徐々に 1 へ近付けるスケジューリングが実用上効果的であると考えられる。どのようなスケジューリングが適切かについては今後の課題である。

本実験ではアルゴリズムが理論通りに収束することを確認するために簡単な問題を取り上げたが，本手法は多くの実問題へも適用可能である。次章では本手法をロボットへ適用する。

3.4 アルゴリズムの特徴のまとめ

本アルゴリズムの特徴を以下にまとめる。

- 各時間ステップでの状態 s_tや割引報酬の期待値 $V_N^{\pi}(s_t)$ 等を明示的に推定するような計算コストのかかる処理が不用であり、実時間処理に向いた方法である。
- 確率的政策$\pi(a, W, X)$ はエージェントの内部変数 W を用いて関数表示されている。そのためニューラルネットやファジィなど任意の関数近似システムを用いることが可能である。これらを用いれば連続値の観測を扱うことも可能である。
- 政策$\pi(a, W, X)$ を行動 a を出力する確率密度関数とすれば、連続値の行動を扱うことができる。
- 本手法を POMDP の環境に適用した場合において、ある観測入力に対して確率的な行動出力へしか収束しない場合には、その観測入力に不完全知覚が存在している。これを利用して不完全知覚を検出することができる。
- 常に環境が変化する場合やマルチエージェント間でのゲームなどの非定常な環境では、時間的に古い過去の環境ほど現在の環境との違いが大きくなっていく。そのため、現在の政策改善に対して過去の経験を利用するにあたり、過去の経験を割り引いて強化することが平均報酬の改善に効果的であると期待できる。また上記のような環境では $\mathrm{Bellman}$ の最適性原理が一般に成り立たない。本手法は $\mathrm{Bellman}$ の最適性原理を利用しないで政策を改善するため、上記のような環境に適している。
- 山登り法の一種と考えられるので局所解へ陥る危険性がある。
- 割引率 γを 1 に近付けることにより、本アルゴリズムによって平均報酬の極大化政策を近似的に得ることができる。
- 存在確率の高い状態ほど優先的に割引報酬を改善していくような合理的な探査戦略を行うメカニズムが含まれている。

第 4 章

ロボットアームの制御問題への適用

4.1 ロボットアームの制御問題

本論文で扱っている POMDP の環境下での強化学習は、第 1 章でも述べたように、故障などやむを得ない事情により限られたセンサの情報やモ- タ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対して最も有効なアプロ- チになると考えられる。本論文ではこのような実問題の一例として、 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問題を考える (図 4.1)。エージェントはある時間間隔でアー

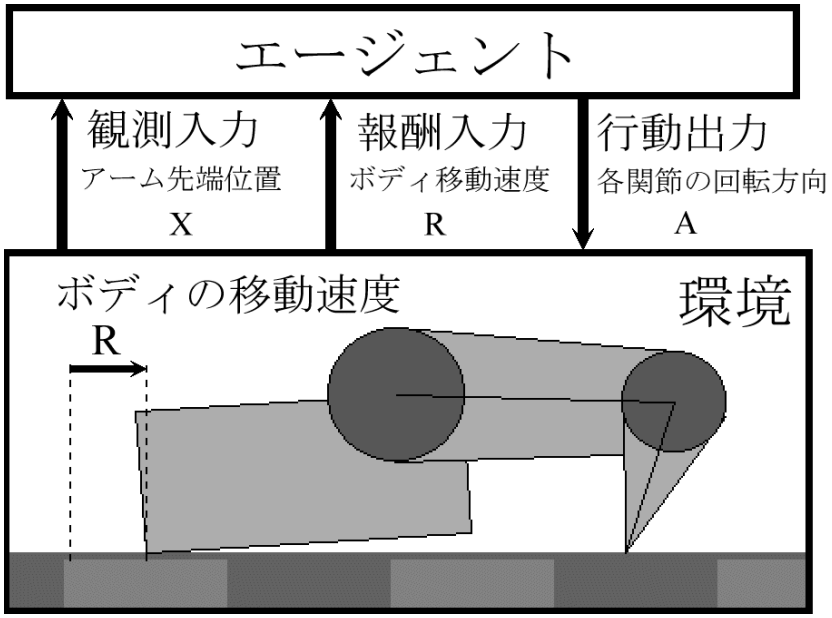

図 4.1: ロボットアームによるほふく前進運動の学習問題

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マルコフ決定過程としてモデル化できる。この学習問題では以下の３点の課題を扱う必要がある。

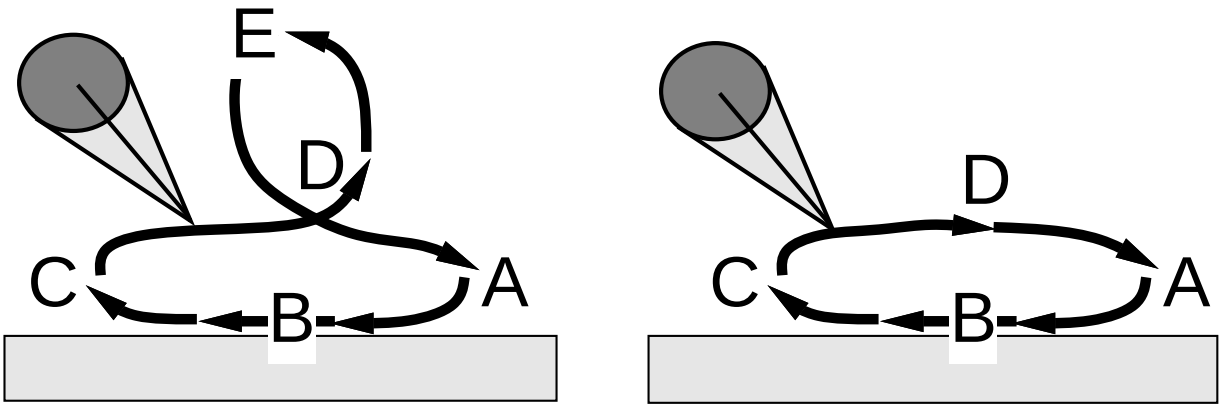

図 4.2: アーム先端の軌跡の例．

(課題１) 報酬の遅れ ： この学習問題ではアーム先端が図 4.2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 $\mathrm{A} \rightarrow \mathrm{B} \rightarrow \mathrm{C}$ 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 $\mathrm{C} \rightarrow \mathrm{D} \rightarrow \mathrm{A}$ 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 4.2の左側のように $\mathrm{C} \rightarrow \mathrm{D} \rightarrow \mathrm{E} \rightarrow \mathrm{D} \rightarrow \mathrm{A}$ 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 ： ロボットのセンサの能力などが不十分なため、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。このように状態観測に不完全性や不確実性がある場合、隠れ状態問題となることが知られている。

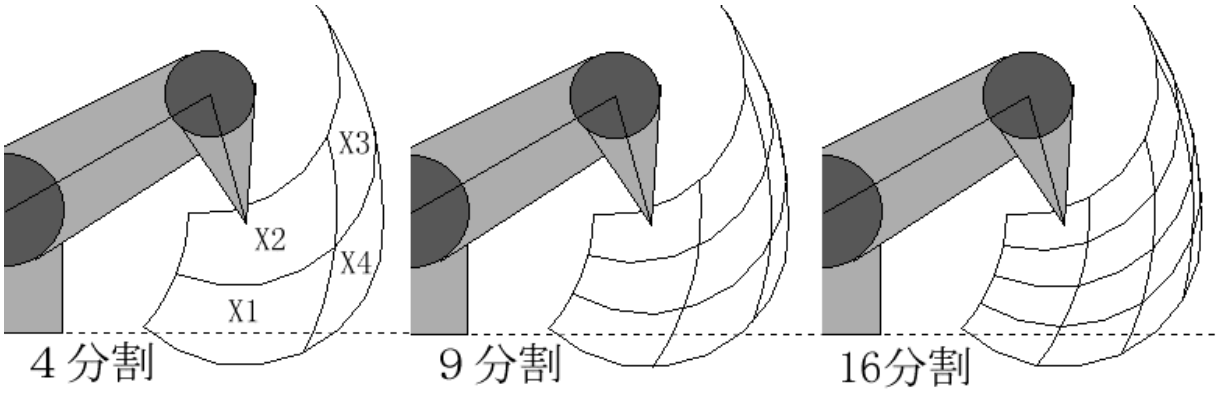

図 4.3: 状態空間の分割

(課題 3) 状態空間の汎化と不完全知覚 ： エー ジェントの計算資源は一般に有限なので、連続で大きな状態空間を持つような環境を扱う場合、観測された状態の空間を適切に汎化する必要がある。例えば連続な状態空間をいくつかの多次元のグリッドに分割して離散化したりニュー ラルネットで適当に補間するなどの方法がある。しかし、このような汎化はエー ジェントのセンサ能力に制限を加えるのと等価であり、隠れ状態問題同様の不完全知覚を誘発する場合がある。

多次元で連続な状態空間を持つマルコフ決定問題を扱う場合には、状態空間を多次元のグリッドに分割してそれぞれのボックスを量子として離散化して扱うのが一般的である。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という 「次元の呪い (curse of dimensionality) 」問題が発生する [Moore et.al 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、しかし不完全知覚による POMDP 問題を引き起こす。強化学習では、エージェントの計算資源に制限があるのに加えて、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については予め知ることはできない。そのため結果的に状態空間の分割が粗くなって、POMDP 問題になることが予想される。

このロボットの問題も状態空間の分割を粗くすると、POMDP 問題になる。例えば図4.3のように、センサ入力である関節の角度空間を 4 等分して離散化する。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という深刻な不完全知覚を生ずる。

状態空間を離散化しないで関節の角度を直接観測できる場合でも、ニューラルネット等による汎化が求められる。このとき、真の value function あるいは政策関数の形状がエージェントの関数近似能力を上回る場合には、何らかの非マルコフ性を生じることがある。

4.2 実験設定

図 4.1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角から水平方向に 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 / 逆回転の指令信号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。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。状態空間は図 4.3のように関節角度の可動範囲の空間を等分割して離散化する。

4.3 エージェントの実装

ロボットに対して図 4.4, 4.5に示す内部構造のエージェントを適用し、それぞれ確率的傾斜法を実装する。

4.3.1 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 4.3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 4.4)。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。

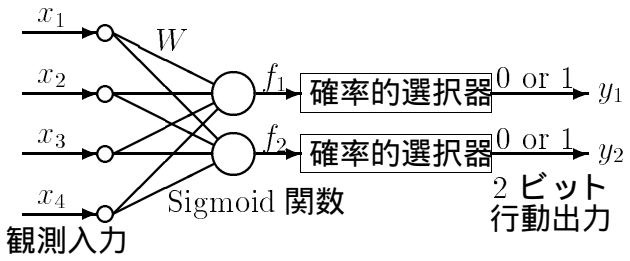

図 4.4: エージェントの内部構造 (離散入力)

4.3.2 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 4.5)。

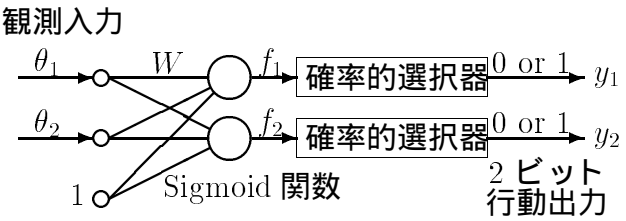

図 4.5: エージェントの内部構造 (連続値入力)

4.3.3 確率的傾斜法の実装

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。

$$f_j = \frac{1}{1 + \exp\left(-\sum_{i=1}^{4} x_i\, w_{ij}\right)} \tag{4.1}$$

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 4.1: 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 4.1 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。

$$
\begin{aligned}
e_{ij}(t) &= \begin{cases} \dfrac{-1}{1-f_j}\dfrac{\partial f_j}{\partial w_{ij}} & , \text{where } y_j = 0, \\ \dfrac{1}{f_j}\dfrac{\partial f_j}{\partial w_{ij}} & , \text{where } y_j = 1. \end{cases} \\
&= \frac{y_j - f_j}{f_j(1-f_j)}\frac{\partial f_j}{\partial w_{ij}} \\
&= x_i(y_j - f_j)\ . \qquad (4.2)
\end{aligned}
$$

よって図 4.4,4.5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 (4.2) を図 3.1の一般形アルゴリズムにあてはめると、図 4.6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 $a_t = (y_1, y_2)$ 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 $(0 \leq \gamma < 1)$ 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、手順 1 へ戻る。

図 4.6: 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。

4.4 実験結果

4.4.1 センサ (角度) 空間をグリッドで分割した場合

図 4.7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 $(4\ \mathrm{cells})$ 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 4.8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 4.8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。

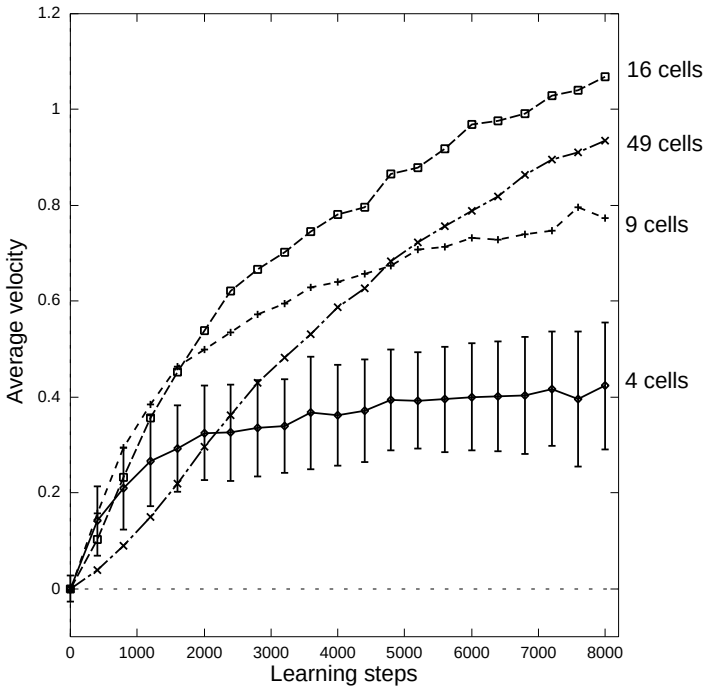

図 4.7: センサ (角度) 空間をグリッドで離散化入力した場合の学習結果。Cells は状態空間の分割個数、横軸は学習ステップ、縦軸は平均速度を表す。

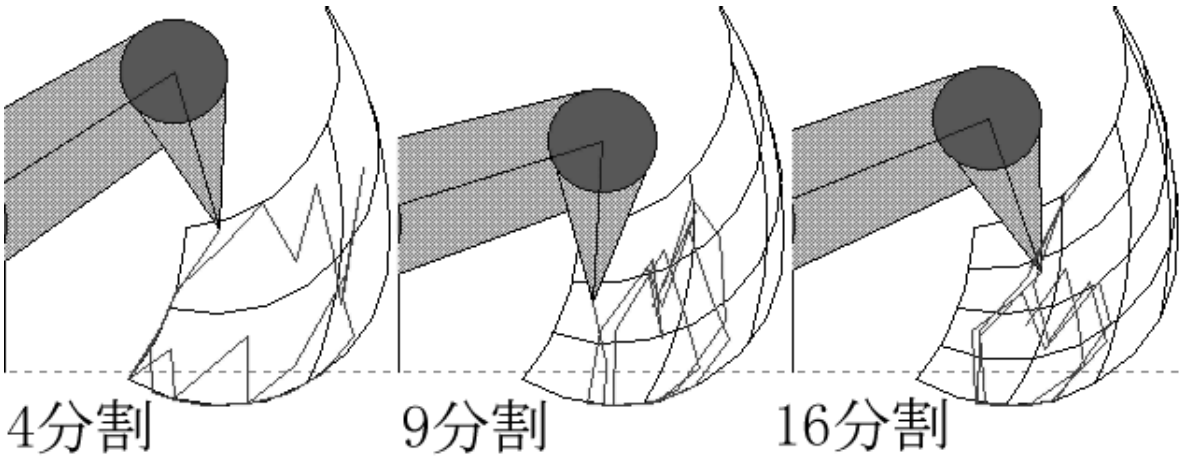

図 4.8: センサ (角度) 空間をグリッドで離散化入力した場合の学習で得たアーム先端の軌跡の一例。

観測入力が離散な場合、Q-learning[Watkins et.al 92] や Jaakkola の手法 [Jaakkola et.al 94] と比較が可能である。図 4.9から図 4.15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 4.9, 4.10 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 4.13, 4.15)。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。

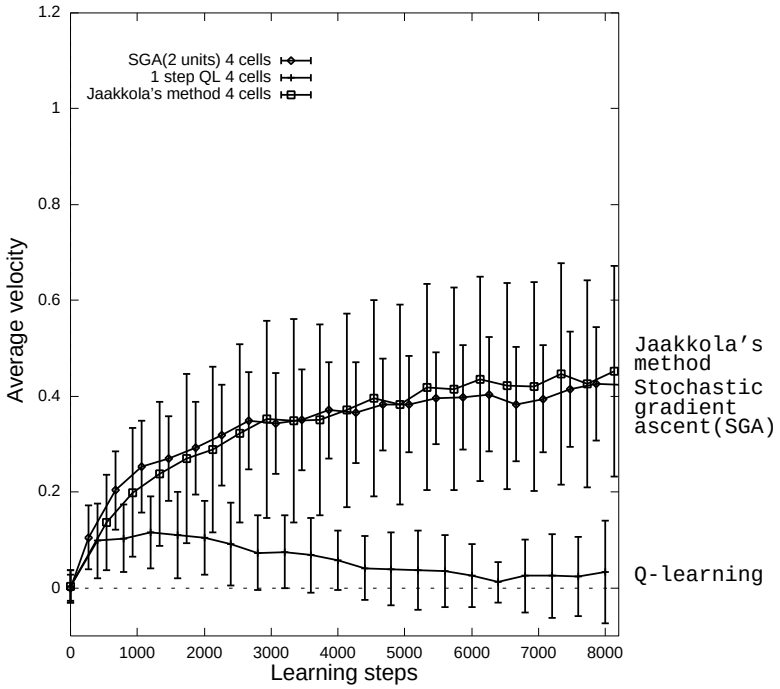

図 4.9: 角度センサの入力空間を 4 分割して離散化した場合の比較。横軸は学習ステップ、縦軸は平均速度を表す。確率的傾斜法は SGA と記した。

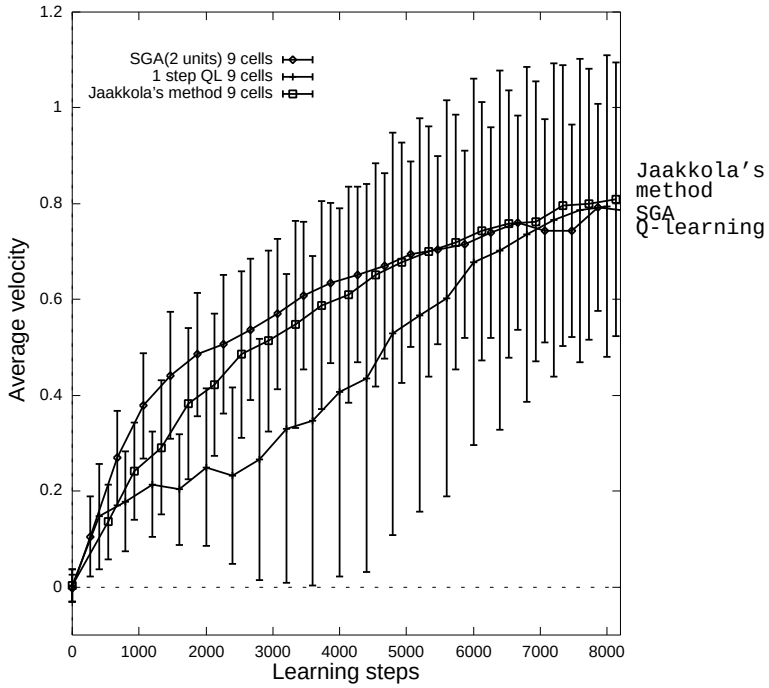

図 4.10: 角度センサの入力空間を 9 分割した場合の比較。

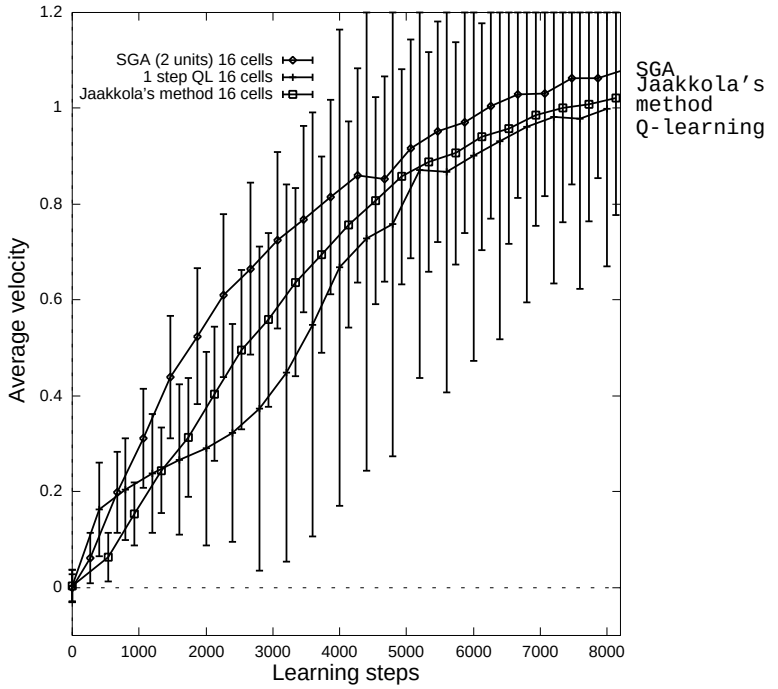

図 4.11: 角度センサの入力空間を 16 分割した場合の比較。

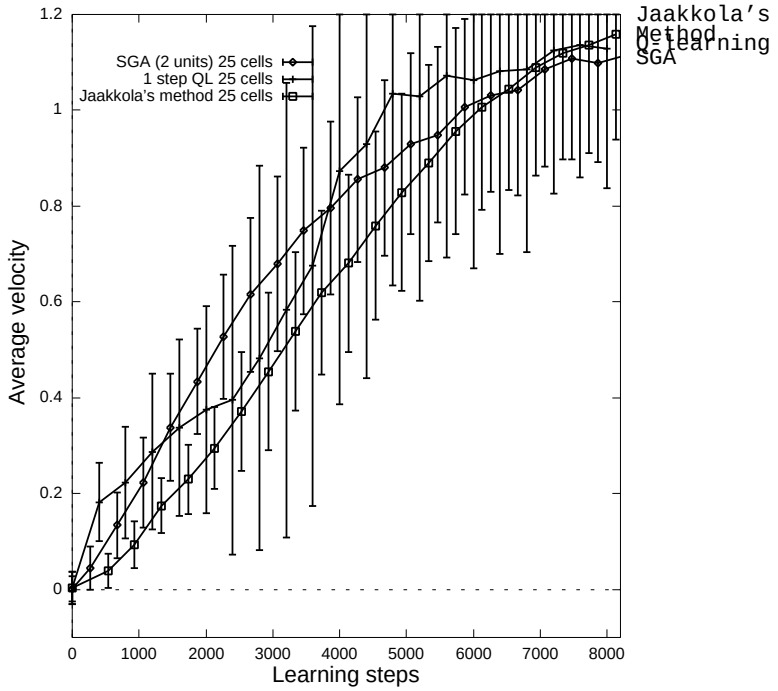

図 4.12: 角度センサの入力空間を 25 分割した場合の比較。

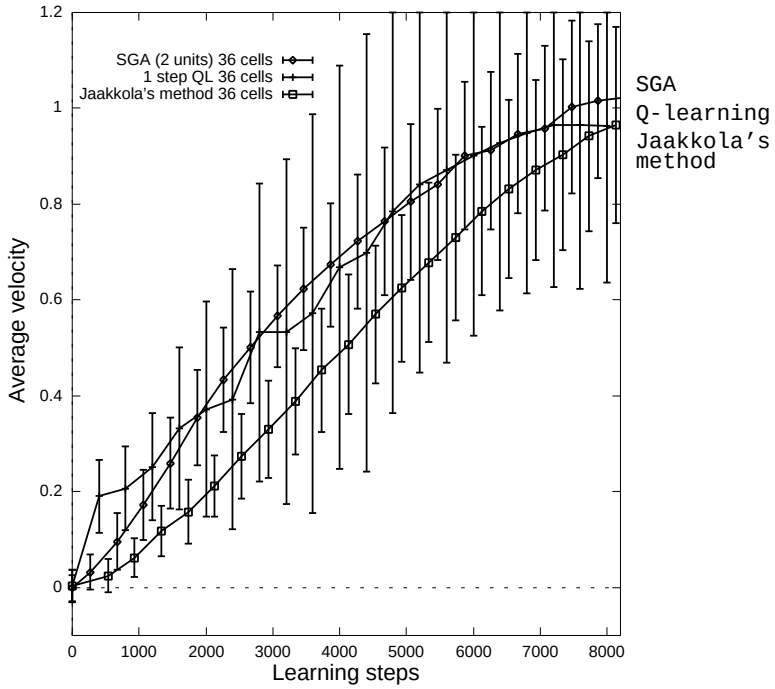

図 4.13: 角度センサの入力空間を 36 分割した場合の比較。

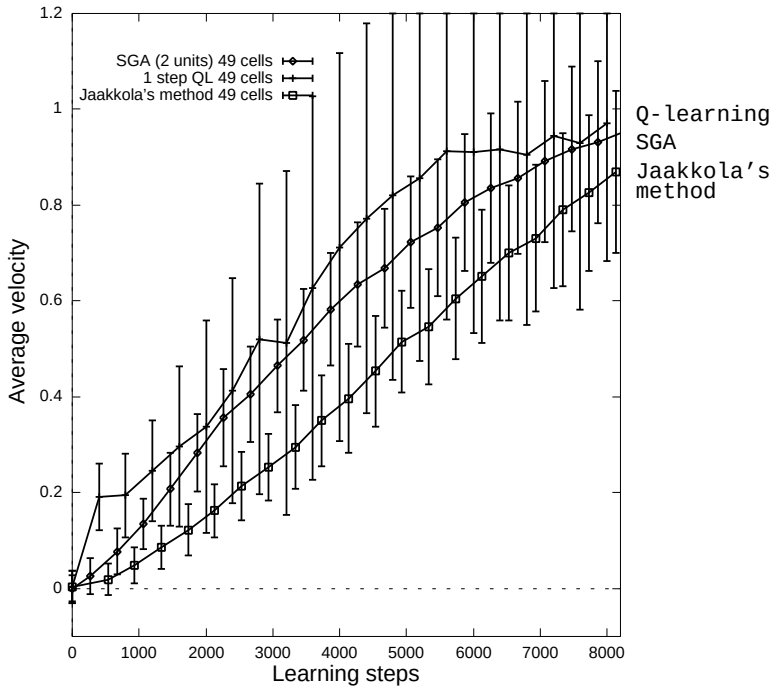

図 4.14: 角度センサの入力空間を 49 分割した場合の比較。

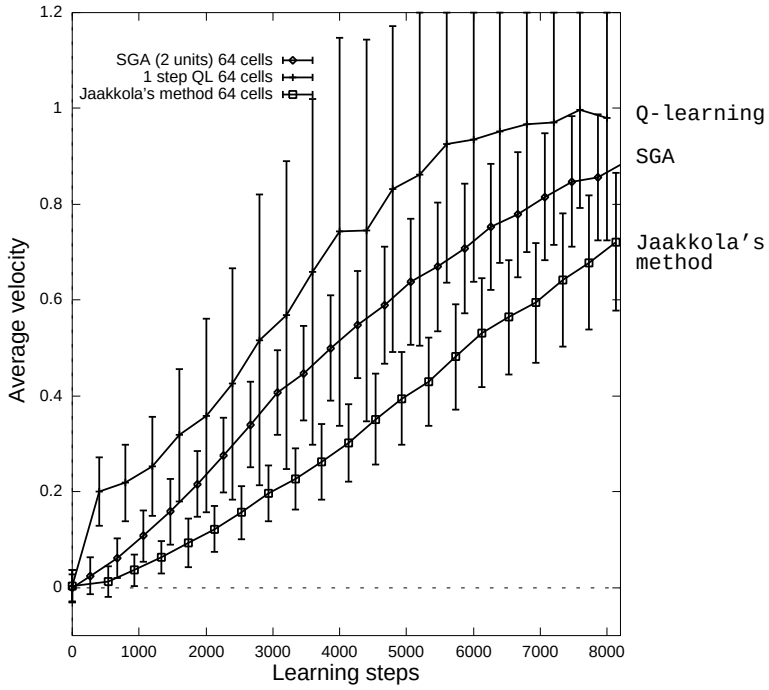

図 4.15: 角度センサの入力空間を 64 分割した場合の比較。Jaakkola の方法は状態空間の増加に対して敏感であることがわかる。

4.4.2 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。4.16 に Q-learning の実装を示す。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。

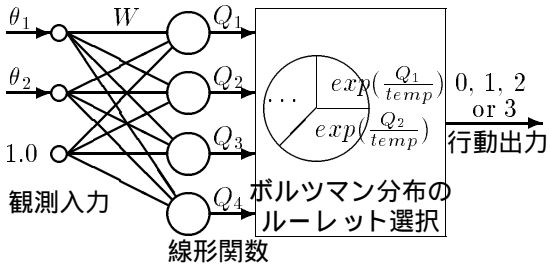

図 4.16: 連続値の観測入力の場合の Q-learning エージェントの実装

図 4.17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。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、Q-learning の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 4.18は学習で得た行動の例を示す。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。

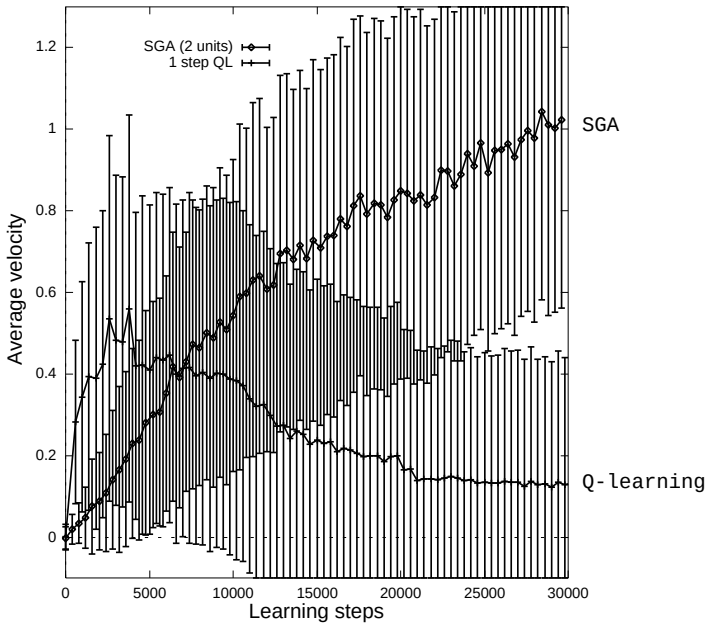

図 4.17: 連続値の観測 (角度) を入力した場合の 100 試行の学習結果。横軸は学習ステップ、縦軸は平均速度を表す。

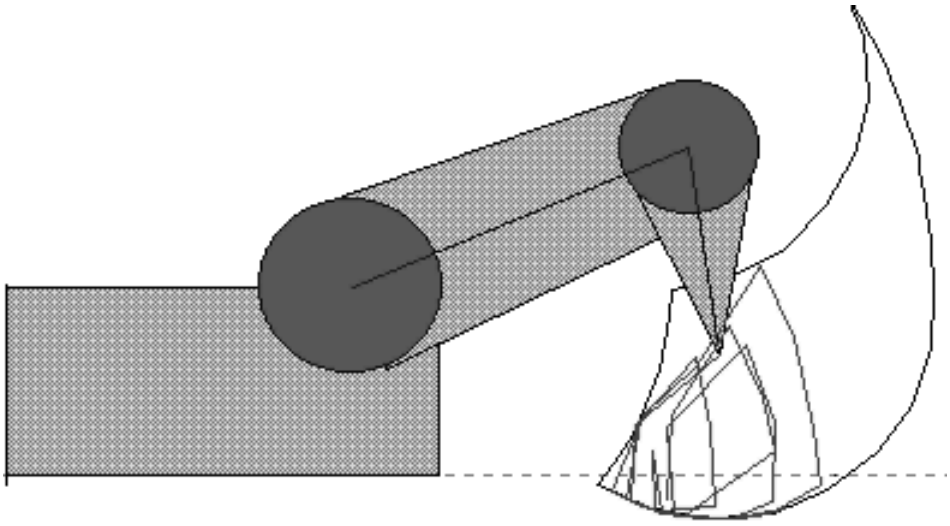

図 4.18: 連続値の観測 (角度) を直接入力して、約 30,000 ステップ学習後に得たアーム先端の軌跡の一例。

4.5 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。

探査戦略の合理性 :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。

確率的政策 :本手法は、他の領域で決定的な行動へと収束していても、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。

第 5 章

結　論

5.1　研究成果のまとめ

本論文では、実世界の強化学習における隠れ状態と関数近似の 2 つの問題を、POMDPs の環境で関数近似を用いる強化学習問題とした。そこで、POMDP の環境において、関数近似されたメモリレスな政策、すなわち観測入力から行動出力確率への単純な写像を形成する強化学習アルゴリズムについて考え、これに基づいて POMDPs の数学的諸性質について述べ、この性質を利用して新しい強化学習アルゴリズムを提案した。

提案手法では、Williams の REINFORCE における適正度 (eligibility) を割引きながら足し合わせた適正度の履歴 (eligibility trace) を用いて政策を逐次的に改善することから、Williams のアルゴリズムの拡張と考えられる。提案手法は各時間ステップでの状態や報酬の期待値等を明示的に推定するような計算コストのかかる処理が不用であり、実時間処理に向いた方法である。また、連続な観測空間や行動空間などへの拡張が容易である。本論文では、提案手法の重みパラメータ更新の平均値が状態の極限分布に重み付けされた割引報酬の期待値の傾斜に等しいことを証明し、確率的傾斜法の一種であることを示した。割引率 γ を 1 に近付けることにより、本アルゴリズムの得る政策を平均報酬の極大化政策に近付けることができる。

さらに、小規模な例題においてシミュレーションを行い、理論的に予想される性質について確認した。ロボットの例題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、提案手法がロバストであることを示した。

5.2 今後の展望

本手法は学習途中など非定常な環境において効率よく学習することも期待されるが、これらの環境における挙動のより詳しい解析やもっと大規模な問題における挙動については今後の課題である。POMDP 環境下での強化学習という枠組は、適応的ファジィ制御の枠組と類似する点が多い。そこで，適応的ファジィ制御に本研究の理論を適用し、メンバーシップ関数の新しいチューニング方法など強化学習に基づく新しい理論を組み立て、設計者の知識を上手に生かすというファジィの利点を強化学習の枠組へ採り入れるなど、実問題への応用を指向した研究も今後の課題である。

本アルゴリズムで用いる計算は、HMM の推定と非常に類似している。これらと本手法との融合による POMDP へのアプローチは、最も興味ある今後の課題である。

付録 A

補題 1 学習定数$\alpha = 0$で政策πが定常な場合，$POMDP$の環境において提案したアルゴリズムによって計算されるΔWは任意の時間ステップ$N > 1$について以下の式が成り立つ．

$$
\begin{aligned}
E\{\Delta w_i(N)\} &= \gamma^N \Big\{ \sum_{s_1} \cdots \sum_{s_N} \frac{\partial}{\partial w_i} \Big(P^\pi(s_0, s_1) \cdots P^\pi(s_{N-1}, s_N) R^\pi(s_N) \Big) \Big\} \\
&\quad + (1-\gamma)\gamma^{N-1} \Big\{ \sum_{s_1} \cdots \sum_{s_N} P^\pi(s_0, s_1) \frac{\partial}{\partial w_i} \Big(P^\pi(s_1, s_2) \cdots P^\pi(s_{N-1}, s_N) R^\pi(s_N) \Big) \Big\} \\
&\quad + \cdots \\
&\quad + (1-\gamma)\gamma^{N-t} \Big\{ \sum_{s_1} \cdots \sum_{s_N} P^\pi(s_0, s_1) \cdots P^\pi(s_{t-1}, s_t) \frac{\partial}{\partial w_i} \Big(P^\pi(s_t, s_{t+1}) \cdots P^\pi(s_{N-1}, s_N) R^\pi(s_N) \Big) \Big\} \\
&\quad + \cdots \\
&\quad + (1-\gamma)\gamma^{1} \Big\{ \sum_{s_1} \cdots \sum_{s_N} P^\pi(s_0, s_1) \cdots P^\pi(s_{N-2}, s_{N-1}) \frac{\partial}{\partial w_i} \Big(P^\pi(s_{N-1}, s_N) R^\pi(s_N) \Big) \Big\} \\
&\quad + (1-\gamma)\gamma^{0} \Big\{ \sum_{s_1} \cdots \sum_{s_N} P^\pi(s_0, s_1) \cdots P^\pi(s_{N-1}, s_N) \frac{\partial}{\partial w_i} R^\pi(s_N) \Big\} , \qquad (5.1)
\end{aligned}
$$

ただし，s_0は初期状態を表し，図 *3.1* のアルゴリズム中の $D_i(t)$ は $t < 0$ において $D_i(t) = 0$ であるとする．

補題 1の証明: 図 3.1 のアルゴリズムより

$$
\begin{aligned}
E\Big\{ \Delta w_i(N) \Big\} &= E\Big\{ (r_N - b)\, D_i(N) \Big\} \\
&= E\Big\{ (r_N - b) \left(e_i(N) + \gamma e_i(N-1) + \gamma^2 e_i(N-2) + \cdots + \gamma^{N-t} e_i(t) + \cdots + \gamma^N e_i(0) \right) \Big\} \\
&= \gamma^N E\Big\{ e_i(0)\, (r_N - b) \Big\} + \gamma^{N-1} E\Big\{ e_i(1)\, (r_N - b) \Big\} + \cdots \\
&\quad + \gamma^{N-t} E\Big\{ e_i(t)\, (r_N - b) \Big\} + \cdots \\
&\quad + \gamma^1 E\Big\{ e_i(N-1)\, (r_N - b) \Big\} + \gamma^0 E\Big\{ e_i(N)\, (r_N - b) \Big\} \qquad (5.2)
\end{aligned}
$$

ここで任意の第 t 項に注目すると，

$$
\begin{aligned}
\gamma^{N-t} E\Big\{ e_i(t)\, (r_N - b) \Big\} &= \gamma^{N-t} \sum_{s_1} \cdots \sum_{s_N} P^\pi(s_0, s_1) \cdots P^\pi(s_{t-1}, s_t) \\
&\quad \times \Big\{ \sum_a \sum_X P(X|s_t)\, \pi(a, W, X) \Big(\frac{\partial}{\partial w_i} \pi(a, W, X) \Big) P^a(s_t, s_{t+1}) \Big\} \\
&\quad \times P^\pi(s_{t+1}, s_{t+2}) \cdots P^\pi(s_{N-1}, s_N)\, R^\pi(s_N) \\
&= \gamma^{N-t} \sum_{s_1} \cdots \sum_{s_N} P^\pi(s_0, s_1) \cdots P^\pi(s_{t-1}, s_t) \\
&\quad \times \Big(\frac{\partial}{\partial w_i} P^\pi(s_t, s_{t+1}) \Big) P^\pi(s_{t+1}, s_{t+2}) \cdots P^\pi(s_{N-1}, s_N)\, R^\pi(s_N) \, .
\end{aligned}
$$

よって

$$
\begin{aligned}
E\Big\{\Delta w_i(N)\Big\} &= \gamma^N\Big\{\sum_{s_1}\cdots\sum_{s_N}\Big(\frac{\partial}{\partial w_i}P^\pi(s_0,s_1)\Big)P^\pi(s_1,s_2)\cdots P^\pi(s_{N-1},s_N)\,R^\pi(s_N)\Big\}\\
&+ \gamma^{N-1}\Big\{\sum_{s_1}\cdots\sum_{s_N}P^\pi(s_0,s_1)\Big(\frac{\partial}{\partial w_i}P^\pi(s_1,s_2)\Big)P^\pi(s_2,s_3)\cdots P^\pi(s_{N-1},s_N)\,R^\pi(s_N)\Big\}\\
&+ \cdots\\
&+ \gamma^{N-t}\Big\{\sum_{s_1}\cdots\sum_{s_N}P^\pi(s_0,s_1)\cdots P^\pi(s_{t-1},s_t)\Big(\frac{\partial}{\partial w_i}P^\pi(s_t,s_{t+1})\Big)\\
&\quad\times P^\pi(s_{t+1},s_{t+2})\cdots P^\pi(s_{N-1},s_N)\,R^\pi(s_N)\Big\}\\
&+ \cdots\\
&+ \gamma^1\Big\{\sum_{s_1}\cdots\sum_{s_N}P^\pi(s_0,s_1)\cdots P^\pi(s_{N-2},s_{N-1})\Big(\frac{\partial}{\partial w_i}P^\pi(s_{N-1},s_N)\Big)R^\pi(s_N)\Big\}\\
&+ \gamma^0\Big\{\sum_{s_1}\cdots\sum_{s_N}P^\pi(s_0,s_1)\cdots P^\pi(s_{N-1},s_N)\Big(\frac{\partial}{\partial w_i}R^\pi(s_N)\Big)\Big\}\ .
\end{aligned}
$$

第 1 項について合成関数の微分の公式に従って整理すると

$$
\begin{aligned}
E\Big\{\Delta w_i(N)\Big\} &= \gamma^N\frac{\partial}{\partial w_i}\Big\{\sum_{s_1}\cdots\sum_{s_N}P^\pi(s_0,s_1)P^\pi(s_1,s_2)\cdots P^\pi(s_{N-1},s_N)\,R^\pi(s_N)\Big\}\\
&+ \gamma^{N-1}(1-\gamma)\Big\{\sum_{s_1}\cdots\sum_{s_N}P^\pi(s_0,s_1)\Big(\frac{\partial}{\partial w_i}P^\pi(s_1,s_2)\Big)P^\pi(s_2,s_3)\cdots P^\pi(s_{N-1},s_N)\,R^\pi(s_N)\Big\}\\
&+ \cdots\\
&+ \gamma^{N-t}(1-\gamma^t)\Big\{\sum_{s_1}\cdots\sum_{s_N}P^\pi(s_0,s_1)\cdots P^\pi(s_{t-1},s_t)\Big(\frac{\partial}{\partial w_i}P^\pi(s_t,s_{t+1})\Big)\\
&\quad\times P^\pi(s_{t+1},s_{t+2})\cdots P^\pi(s_{N-1},s_N)\,R^\pi(s_N)\Big\}\\
&+ \cdots\\
&+ \gamma^1(1-\gamma^{N-1})\Big\{\sum_{s_1}\cdots\sum_{s_N}P^\pi(s_0,s_1)\cdots P^\pi(s_{N-2},s_{N-1})\Big(\frac{\partial}{\partial w_i}P^\pi(s_{N-1},s_N)\Big)R^\pi(s_N)\Big\}\\
&+ \gamma^0(1-\gamma^N)\Big\{\sum_{s_1}\cdots\sum_{s_N}P^\pi(s_0,s_1)\cdots P^\pi(s_{N-1},s_N)\Big(\frac{\partial}{\partial w_i}R^\pi(s_N)\Big)\Big\}\ .
\end{aligned}
$$

以下同様に第 2 項以降について順次整理していくと，ただちに (5.1) 式を得る． ■

補題 2 学習定数$\alpha = 0$ で政策πが定常な場合，$POMDP$ の環境において提案したアルゴリズムによって計算されるΔWは任意の時間ステップ $N \geq 1$ について以下の式を満たす．

$$
E\Big\{\sum_{t=0}^{N}\Delta w_i(t)\Big\} = \Big(\frac{\partial}{\partial w_i}V_N^\pi(s_0)\Big) + (1-\gamma)\Big\{\sum_{t=1}^{N}\sum_{s_1}\cdots\sum_{s_t}P^\pi(s_0,s_1)\cdots P^\pi(s_{t-1},s_t)\Big(\frac{\partial}{\partial w_i}V_{N-t}^\pi(s_t)\Big)\Big\} \tag{5.3}
$$

ただし，s_0は初期状態を表す．図 3.1 のアルゴリズム中の $D_i(t)$ は $t<0$ において $D_i(t)=0$ であるとする．

補題 2の証明: $N = 1$ のとき

$$
E\Big\{\sum_{t=0}^{1}\Delta w_i(t)\Big\} = E\Big\{(r_0-b)\,e_i(0) + (r_1-b)\,\big(e_i(1)+\gamma e_i(0)\big)\Big\}
$$

$$
\begin{aligned}
&= E\Big\{(r_0-b)\,e_i(0)\Big\} + E\Big\{(r_1-b)\,e_i(1)\Big\} + E\Big\{(r_1-b)\,\gamma e_i(0)\Big\} \\
&= \Big\{\sum_a \sum_X P(X|s_0)\pi(a,W,X)\Big(\frac{\partial}{\partial w_i}\pi(a,W,X)\Big)(R^a(s_0)-b)\Big\} \\
&\quad +\Big\{\sum_{s_1} P^\pi(s_0,s_1)\sum_a\sum_X P(X|s_1)\pi(a,w,X)\Big(\frac{\partial}{\partial w_i}\pi(a,W,X)\Big)(R^a(s_1)-b)\Big\} \\
&\quad +\gamma\Big\{\sum_{s_1}\sum_a\sum_X P(X|s_0)\pi(a,W,X)\Big(\frac{\partial}{\partial w_i}\pi(a,W,X)\Big)P^a(s_0,s_1)\,(R^\pi(s_1)-b)\Big\} \\
&= \Big(\frac{\partial}{\partial w_i}R^\pi(s_0)\Big) + \Big\{\sum_{s_1} P^\pi(s_0,s_1)\Big(\frac{\partial}{\partial w_i}R^\pi(s_1)\Big)\Big\} + \gamma\Big\{\sum_{s_1}\Big(\frac{\partial}{\partial w_i}P^\pi(s_0,s_1)\Big)R^\pi(s_1)\Big\} \\
&= \Big\{\frac{\partial}{\partial w_i}\Big(R^\pi(s_0)+\gamma\sum_{s_1}P^\pi(s_0,s_1)R^\pi(s_1)\Big)\Big\} + (1-\gamma)\Big\{\sum_{s_1}P^\pi(s_0,s_1)\Big(\frac{\partial}{\partial w_i}R^\pi(s_1)\Big)\Big\} \\
&= \Big(\frac{\partial}{\partial w_i}V_1^\pi(s_0)\Big) + (1-\gamma)\Big\{\sum_{s_1}P^\pi(s_0,s_1)\Big(\frac{\partial}{\partial w_i}V_0^\pi(s_1)\Big)\Big\}\,.
\end{aligned}
$$

よって $N=1$ のとき (5.3) 式は成り立つ．
$N=k$のとき (5.3) 式が成り立つと仮定すると，$N=k+1$ のとき

$$
E\Big\{\sum_{t=0}^{k+1}\Delta w_i(t)\Big\} \;=\; E\Big\{\sum_{t=0}^{k}\Delta w_i(t)\Big\} + E\Big\{\Delta w_i(k+1)\Big\}\,.
$$

ここで $N=k$の (5.3) 式と $N=k+1$ とした (5.1) 式を代入して整理すると，

$$
\begin{aligned}
E\Big\{\sum_{t=0}^{k+1}\Delta w_i(t)\Big\} &= \Big(\frac{\partial}{\partial w_i}V_k^\pi(s_0)\Big) + \gamma^{k+1}\Big\{\sum_{s_1}\cdots\sum_{s_{k+1}}\frac{\partial}{\partial w_i}\Big(P^\pi(s_0,s_1)\cdots P^\pi(s_k,s_{k+1})R^\pi(s_{k+1})\Big)\Big\} \\
&\quad +(1-\gamma)\Big\{\sum_{t=1}^{k}\sum_{s_1}\cdots\sum_{s_t}P^\pi(s_0,s_1)\cdots P^\pi(s_{t-1},s_t)\Big(\frac{\partial}{\partial w_i}V_{k-t}^\pi(s_t)\Big)\Big\} \\
&\quad +(1-\gamma)\Big\{\sum_{t=1}^{k}\gamma^{k-t}\sum_{s_1}\cdots\sum_{s_t}P^\pi(s_0,s_1)\cdots P^\pi(s_{t-1},s_t)\frac{\partial}{\partial w_i}\Big(P^\pi(s_t,s_{t+1})\cdots \\
&\qquad P^\pi(s_k,s_{k+1})R^\pi(s_{k+1})\Big)\Big\} \\
&\quad +(1-\gamma)\gamma^0\Big\{\sum_{s_1}\cdots\sum_{s_{k+1}}P^\pi(s_0,s_1)\cdots P^\pi(s_k,s_{k+1})\Big(\frac{\partial}{\partial w_i}R^\pi(s_{k+1})\Big)\Big\} \\
&= \Big(\frac{\partial}{\partial w_i}V_{k+1}^\pi(s_0)\Big) + (1-\gamma)\Big\{\sum_{t=1}^{k+1}\sum_{s_1}\cdots\sum_{s_t}P^\pi(s_0,s_1)\cdots P^\pi(s_{t-1},s_t)\Big(\frac{\partial}{\partial w_i}V_{k-t}^\pi(s_t)\Big)\Big\}\,.
\end{aligned}
$$

よって $N=k+1$ のとき (5.3) 式は成り立つ．
数学的帰納法により，全ての自然数 Nについて (5.3) 式は成り立つ． ■

以下では証明のため以下の値を定義する．
全ての時間ステップ $t\geq 0$ において報酬の絶対値 $|r_t|$ が上界を持つとき，$\sup|r_t| = R_{max}$ と表す．
全ての $a\in\mathcal{A}$, $X\in\mathcal{X}$, W において $|\frac{\partial}{\partial w_i}\ln\pi(a,W,X)|$ が上界を持つとき，$\sup|\frac{\partial}{\partial w_i}\ln\pi(a,W,X)| = e_{max}$ と表す．

Fact 1 全ての状態 s, 政策πにおいて

$$
|V_\infty^\pi(s)| \;\leq\; \frac{R_{max}}{1-\gamma}\,. \tag{5.4}
$$

Fact 1 の証明: (2.15) 式より，明らかに

$$|V_\infty^\pi(s)| \ \leq \ \sum_{t=0}^{\infty} \gamma^t R_{max} \ .$$

よって直ちに (5.4) 式を得る． ■

Fact 2 全ての状態 s, 政策πにおいて

$$\left| \frac{\partial}{\partial w_i} R^\pi(s) \right| \ \leq \ e_{max} \, R_{max} \ . \tag{5.5}$$

Fact 2 の証明: (2.4) 式より

$$\begin{aligned} \frac{\partial}{\partial w_i} R^\pi(s) &= \frac{\partial}{\partial w_i} \Big\{ \sum_{a \in \mathcal{A}} \sum_{X \in \mathcal{X}} P(X|s) \, \pi(a, W, X) \, R^a(s) \Big\} \\ &= \sum_{a \in \mathcal{A}} \sum_{X \in \mathcal{X}} P(X|s) \, \Big(\frac{\partial}{\partial w_i} \pi(a, W, X) \Big) \, R^a(s) \\ &= \sum_{a \in \mathcal{A}} \sum_{X \in \mathcal{X}} P(X|s) \, \pi(a, W, X) \Big(\frac{\partial}{\partial w_i} \ln \pi(a, W, X) \Big) \, R^a(s) \ . \end{aligned} \tag{5.6}$$

ここで，(5.6) 式は明らかに

$$\left| \sum_{a \in \mathcal{A}} \sum_{X \in \mathcal{X}} P(X|s) \, \pi(a, W, X) \Big(\frac{\partial}{\partial w_i} \ln \pi(a, W, X) \Big) \, R^a(s) \right| \leq \sum_{a \in \mathcal{A}} \sum_{X \in \mathcal{X}} P(X|s) \, \pi(a, W, X) \, e_{max} \, R_{max} \ . \tag{5.7}$$

e_{max} と R_{max} は定数，$P(X|s)$ と $\pi(a, W, X)$ は確率なので (5.7) 式の不等式の右辺は (5.5) 式の右辺に等しい．■

Fact 3 全ての状態 s, s', 政策πにおいて

$$\left| \frac{\partial}{\partial w_i} P^\pi(s, s') \right| \ \leq \ e_{max} \ . \tag{5.8}$$

Fact 3 の証明: (2.2) 式より

$$\begin{aligned} \frac{\partial}{\partial w_i} P^\pi(s, s') &= \frac{\partial}{\partial w_i} \Big\{ \sum_{a \in \mathcal{A}} \sum_{X \in \mathcal{X}} P(X|s) \, \pi(a, W, X) \, P^a(s, s') \Big\} \\ &= \sum_{a \in \mathcal{A}} \sum_{X \in \mathcal{X}} P(X|s) \, \Big(\frac{\partial}{\partial w_i} \pi(a, W, X) \Big) \, P^a(s, s') \\ &= \sum_{a \in \mathcal{A}} \sum_{X \in \mathcal{X}} P(X|s) \, \pi(a, W, X) \Big(\frac{\partial}{\partial w_i} \ln \pi(a, W, X) \Big) \, P^a(s, s') \ . \end{aligned} \tag{5.9}$$

ここで，(5.9) 式は明らかに

$$\left| \sum_{a \in \mathcal{A}} \sum_{X \in \mathcal{X}} P(X|s) \, \pi(a, W, X) \Big(\frac{\partial}{\partial w_i} \ln \pi(a, W, X) \Big) \, P^a(s, s') \right| \leq \sum_{a \in \mathcal{A}} \sum_{X \in \mathcal{X}} P(X|s) \, \pi(a, W, X) \, e_{max} \, P^a(s, s') \ . \tag{5.10}$$

e_{max} は定数，$P(X|s)$ と $\pi(a, W, X)$ と $P^a(s, s')$ は確率である．従って (5.10) 式の不等式の右辺を整理すると (5.8) 式の右辺に等しい． ■

Fact 4 全ての状態 s, 政策πにおいて

$$\left|\frac{\partial}{\partial w_i}V_N^\pi(s)\right| \;\leq\; \frac{1-(N+2)\gamma^{N+1}+(N+1)\gamma^{N+2}}{(1-\gamma)^2}\,e_{max}\,R_{max}\ . \tag{5.11}$$

Fact 4 の証明: (2.15) 式を w_iで微分するものとして，任意の第 $t+1$ 項 $(0\leq t\leq N)$ について考えると，

$$\begin{aligned}
&\frac{\partial}{\partial w_i}\gamma^t\Big\{\sum_{s_1\in\mathcal{S}}\cdots\sum_{s_t\in\mathcal{S}}P^\pi(s_0,s_1)\cdots P^\pi(s_{t-1},s_t)\,R^\pi(s_t)\Big\}\\
&=\;\gamma^t\Big\{\sum_{s_1\in\mathcal{S}}\cdots\sum_{s_t\in\mathcal{S}}P^\pi(s_0,s_1)\cdots P^\pi(s_{t-1},s_t)\frac{\partial}{\partial w_i}\,R^\pi(s_t)\Big\}\\
&\quad+\gamma^t\Big\{\sum_{s_1\in\mathcal{S}}\cdots\sum_{s_t\in\mathcal{S}}\Big(\frac{\partial}{\partial w_i}\,P^\pi(s_0,s_1)\cdots P^\pi(s_{t-1},s_t)\Big)\,R^\pi(s_t)\Big\}\ .
\end{aligned}\tag{5.12}$$

ここで，(5.12) 式に注目すると

$$\left|\gamma^t\Big\{\sum_{s_1\in\mathcal{S}}\cdots\sum_{s_t\in\mathcal{S}}P^\pi(s_0,s_1)\cdots P^\pi(s_{t-1},s_t)\frac{\partial}{\partial w_i}\,R^\pi(s_t)\Big\}\right| \;\leq\; \gamma^t\,e_{max}\,R_{max}\ , \tag{5.13}$$

$$\left|\gamma^t\Big\{\sum_{s_1\in\mathcal{S}}\cdots\sum_{s_t\in\mathcal{S}}\Big(\frac{\partial}{\partial w_i}\,P^\pi(s_0,s_1)\cdots P^\pi(s_{t-1},s_t)\Big)\,R^\pi(s_t)\Big\}\right| \;\leq\; t\,\gamma^t\,e_{max}\,R_{max}\ . \tag{5.14}$$

よって (5.12)，(5.13)，(5.14) 式より

$$\left|\frac{\partial}{\partial w_i}\gamma^t\Big\{\sum_{s_1\in\mathcal{S}}\cdots\sum_{s_t\in\mathcal{S}}P^\pi(s_0,s_1)\cdots P^\pi(s_{t-1},s_t)\,R^\pi(s_t)\Big\}\right|\leq\gamma^t(t+1)\,e_{max}\,R_{max}\ .$$

これを (2.15) 式の全ての項について考えると，

$$\left|\frac{\partial}{\partial w_i}V_N^\pi(s)\right|\leq\sum_{t=0}^{N}\gamma^t(t+1)\,e_{max}\,R_{max}\ . \tag{5.15}$$

級数の公式を用いると (5.14) 式の不等式の右辺は (5.11) 式の右辺に等しい．

■

Fact 5 全ての状態 s, 政策πにおいて

$$\left|\frac{\partial}{\partial w_i}\Big(V_\infty^\pi(s)-V_N^\pi(s)\Big)\right| \;\leq\; \frac{\gamma^{N+1}}{1-\gamma}\Big(N+1+\frac{1}{1-\gamma}\Big)\,e_{max}\,R_{max}\ . \tag{5.16}$$

Fact 5 の証明: (2.15) 式より

$$\begin{aligned}
\frac{\partial}{\partial w_i}\Big(V_\infty^\pi(s)-V_N^\pi(s)\Big) &= \frac{\partial}{\partial w_i}\Big(\gamma^{N+1}\sum_{s_1}\cdots\sum_{s_{N+1}}P^\pi(s,s_1)\cdots P^\pi(s_N,s_{N+1})V_\infty^\pi(s_{N+1})\Big)\\
&=\gamma^{N+1}\sum_{s_1}\cdots\sum_{s_{N+1}}\frac{\partial}{\partial w_i}\Big(P^\pi(s,s_1)\cdots P^\pi(s_N,s_{N+1})\Big)V_\infty^\pi(s_{N+1})\\
&\quad+\gamma^{N+1}\sum_{s_1}\cdots\sum_{s_{N+1}}P^\pi(s,s_1)\cdots P^\pi(s_N,s_{N+1})\frac{\partial}{\partial w_i}V_\infty^\pi(s_{N+1})
\end{aligned}\tag{5.17}$$

ここで Fact 4 より，

$$\lim_{N\to\infty} \frac{1-(N+2)\gamma^{N+1}+(N+1)\gamma^{N+2}}{(1-\gamma)^2}\, e_{max}\, R_{max} = \frac{1}{(1-\gamma)^2}\, e_{max}\, R_{max}\ ,$$

よって

$$\left|\frac{\partial}{\partial w_i} V_\infty^\pi(s)\right| \ \leq\ \frac{1}{(1-\gamma)^2}\, e_{max}\, R_{max}\ . \tag{5.18}$$

ここで Fact1，Fact3，(5.18) 式より

$$\left|\gamma^{N+1}\sum_{s_1}\cdots\sum_{s_{N+1}}\frac{\partial}{\partial w_i}\Big(P^\pi(s,s_1)\cdots P^\pi(s_N,s_{N+1})\Big)V_\infty^\pi(s_{N+1})\right| \ \leq\ \gamma^{N+1}(N+1)\, e_{max}\,\frac{R_{max}}{1-\gamma} \tag{5.19}$$

$$\left|\gamma^{N+1}\sum_{s_1}\cdots\sum_{s_{N+1}}P^\pi(s,s_1)\cdots P^\pi(s_N,s_{N+1})\frac{\partial}{\partial w_i}V_\infty^\pi(s_{N+1})\right| \ \leq\ \gamma^{N+1}\,\frac{e_{max}\,R_{max}}{(1-\gamma)^2}\ . \tag{5.20}$$

よって (5.17)，(5.19)，(5.20) 式より (5.16) 式を得る． ■

補題 3 提案したアルゴリズムにおいて政策πが定常な場合，以下の式を満たす Nが存在する。

$$\left|\frac{E\Big\{\sum_{t=0}^N \frac{\Delta w_i(t)}{1-\gamma}\Big\} - \sum_{t=0}^N\sum_{s_1}\cdots\sum_{s_t}P^\pi(s_0,s_1)\cdots P^\pi(s_{t-1},s_t)\frac{\partial}{\partial w_i}V_\infty^\pi(s_t)}{N}\right| \ <\ \epsilon_1\ . \tag{5.21}$$

ただし図 *3.1* のアルゴリズム中の $D_i(t)$ は $t<0$ において $D_i(t)=0$ であるとする．

補題 3の証明: (5.21) 式の左辺の分子は補題 2より

$$\begin{aligned}
&E\Big\{\sum_{t=0}^N\frac{\Delta w_i(t)}{1-\gamma}\Big\} - \sum_{t=0}^N\sum_{s_1}\cdots\sum_{s_t}P^\pi(s_0,s_1)\cdots P^\pi(s_{t-1},s_t)\frac{\partial}{\partial w_i}V_\infty^\pi(s_t)\\
&= \frac{1}{1-\gamma}\frac{\partial}{\partial w_i}V_N^\pi(s_0) + \Big\{\sum_{t=1}^N\sum_{s_1}\cdots\sum_{s_t}P^\pi(s_0,s_1)\cdots P^\pi(s_{t-1},s_t)\Big(\frac{\partial}{\partial w_i}V_{N-t}^\pi(s_t)\Big)\Big\}\\
&\quad - \Big\{\sum_{t=1}^N\sum_{s_1}\cdots\sum_{s_t}P^\pi(s_0,s_1)\cdots P^\pi(s_{t-1},s_t)\Big(\frac{\partial}{\partial w_i}V_\infty^\pi(s_t)\Big)\Big\}\\
&= \frac{1}{1-\gamma}\frac{\partial}{\partial w_i}V_N^\pi(s_0) - \Big\{\sum_{t=1}^N\sum_{s_1}\cdots\sum_{s_t}P^\pi(s_0,s_1)\cdots P^\pi(s_{t-1},s_t)\frac{\partial}{\partial w_i}\Big(V_\infty^\pi(s_t)-V_{N-t}^\pi(s_t)\Big)\Big\}
\end{aligned} \tag{5.22}$$

ここで Fact4，Fact5より

$$\begin{aligned}
&\left|\frac{1}{1-\gamma}\frac{\partial}{\partial w_i}V_N^\pi(s_0)\right| + \left|\Big\{\sum_{t=1}^N\sum_{s_1}\cdots\sum_{s_t}P^\pi(s_0,s_1)\cdots P^\pi(s_{t-1},s_t)\frac{\partial}{\partial w_i}\Big(V_\infty^\pi(s_t)-V_{N-t}^\pi(s_t)\Big)\Big\}\right|\\
&\leq \frac{1-(N+2)\gamma^{N+1}+(N+1)\gamma^{N+2}}{(1-\gamma)^3}\, e_{max}\, R_{max} + \sum_{t=1}^N\Big\{\frac{\gamma^{(N-t)+1}}{1-\gamma}\big((N-t)+1+\frac{1}{1-\gamma}\big)\, e_{max}\, R_{max}\Big\}
\end{aligned} \tag{5.23}$$

ここで，

$$\frac{1-(N+2)\gamma^{N+1}+(N+1)\gamma^{N+2}}{(1-\gamma)^3}\, e_{max}\, R_{max} + \sum_{t=1}^N\Big\{\frac{\gamma^{(N-t)+1}}{1-\gamma}\big((N-t)+1+\frac{1}{1-\gamma}\big)\, e_{max}\, R_{max}\Big\}$$

$$
\begin{aligned}
&= \Big(\frac{1-(N+2)\gamma^{N+1}+(N+1)\gamma^{N+2}}{(1-\gamma)^3}+\frac{1}{1-\gamma}\Big(\sum_{t=1}^{N} t\,\gamma^t\Big)+\frac{1}{(1-\gamma)^2}\Big(\sum_{t=1}^{N}\gamma^t\Big)\Big)\, e_{max}\, R_{max} \\
&= \Big(\frac{1-(N+2)\gamma^{N+1}+(N+1)\gamma^{N+2}}{(1-\gamma)^3}+\gamma\frac{1-(N+1)\gamma^{N}+N\gamma^{N+1}}{(1-\gamma)^3}+\gamma\frac{1-\gamma^N}{(1-\gamma)^3}\Big)\, e_{max}\, R_{max} \\
&\le \frac{1+2\gamma}{(1-\gamma)^3}\, e_{max}\, R_{max}\ . \qquad (5.24)
\end{aligned}
$$

よって (5.21) 式の左辺の分子は (5.22)，(5.23)，(5.24) 式より

$$
\Big|E\Big\{\sum_{t=0}^{N}\frac{\Delta w_i(t)}{1-\gamma}\Big\}-\sum_{t=0}^{N}\sum_{s_1}\cdots\sum_{s_t}P^\pi(s_0,s_1)\cdots P^\pi(s_{t-1},s_t)\frac{\partial}{\partial w_i}V_\infty^\pi(s_t)\Big| \le \frac{1+2\gamma}{(1-\gamma)^3}\, e_{max}\, R_{max}\ . \qquad (5.25)
$$

(5.25) 式の右辺は正の定数である．従って

$$
\frac{1+2\gamma}{(1-\gamma)^3}\, e_{max}\, R_{max} \ < \ \epsilon_1\, m\ ,
$$

となる自然数 m が存在し，$N > m$ の全ての Nについて (5.21) 式が成り立つ． ■

Fact 6 政策πが定常な場合，

$$
\lim_{N\to\infty}\Big\{\sum_{s_1}\cdots\sum_{s_N}P^\pi(s_0,s_1)\cdots P^\pi(s_{N-1},s_N)\Big(\frac{\partial}{\partial w_i}V_\infty^\pi(s_N)\Big)\Big\} \ = \ \sum_s U^\pi(s)\frac{\partial}{\partial w_i}V_\infty^\pi(s)\ . \qquad (5.26)
$$

Fact 6 の証明: 状態 s_N を s と表すと，

$$
\begin{aligned}
&\lim_{N\to\infty}\Big\{\sum_{s_1}\cdots\sum_{s_N}P^\pi(s_0,s_1)\cdots P^\pi(s_{N-1},s_N)\Big(\frac{\partial}{\partial w_i}V_\infty^\pi(s_N)\Big)\Big\} \\
&= \lim_{N\to\infty}\sum_s\Big\{\sum_{s_1}\cdots\sum_{s_{N-1}}P^\pi(s_0,s_1)\cdots P^\pi(s_{N-1},s)\Big(\frac{\partial}{\partial w_i}V_\infty^\pi(s)\Big)\Big\} \\
&= \sum_s\lim_{N\to\infty}\Big\{\sum_{s_1}\cdots\sum_{s_{N-1}}P^\pi(s_0,s_1)\cdots P^\pi(s_{N-1},s)\Big\}\Big(\frac{\partial}{\partial w_i}V_\infty^\pi(s)\Big)\ . \qquad (5.27)
\end{aligned}
$$

ここで，エルゴート性を有するマルコフ過程なので

$$
\lim_{N\to\infty}\sum_{s_1}\cdots\sum_{s_{N-1}}P^\pi(s_0,s_1)\cdots P^\pi(s_{N-1},s)=U^\pi(s)\ . \qquad (5.28)
$$

である．よって (5.27)，(5.28) 式より直ちに Fact 6を得る． ■

Fact 7 政策πが定常な場合，

$$
\lim_{N\to\infty}\frac{\sum_{t=0}^{N}\Big\{\sum_{s_1}\cdots\sum_{s_t}P^\pi(s_0,s_1)\cdots P^\pi(s_{t-1},s_t)\Big(\frac{\partial}{\partial w_i}V_\infty^\pi(s_t)\Big)\Big\}}{N} \ = \ \sum_s U^\pi(s)\frac{\partial}{\partial w_i}V_\infty^\pi(s)\ . \qquad (5.29)
$$

Fact 7 の証明: ある数列 A_n が $\lim_{n\to\infty} A_n = A$ のとき，

$$\lim_{n\to\infty} \frac{\sum_{k=0}^{n} A_k}{n} = A \ .$$

である．ここで Fact6より

$$\begin{aligned} A_n &= \sum_{s_1} \cdots \sum_{s_N} P^\pi(s_0, s_1) \cdots P^\pi(s_{N-1}, s_N) \Big(\frac{\partial}{\partial w_i} V_\infty^\pi(s_N) \Big) \ , \\ A &= \sum_{s} U^\pi(s) \frac{\partial}{\partial w_i} V_\infty^\pi(s) \ . \end{aligned}$$

とおけば直ちに得る． ■

定理 2の証明: 中心値極限定理より

$$\lim_{N\to\infty} \frac{1}{(1-\gamma)N} \sum_{t=0}^{N-1} \Delta w_i(t) = \lim_{N\to\infty} \frac{1}{(1-\gamma)N} E\left\{ \sum_{t=0}^{N-1} \Delta w_i(t) \right\} \ . \tag{5.30}$$

ここで補題 3より

$$\lim_{N\to\infty} \frac{1}{(1-\gamma)N} E\left\{ \sum_{t=0}^{N-1} \Delta w_i(t) \right\} = \lim_{N\to\infty} \frac{\sum_{t=0}^{N} \sum_{s_1} \cdots \sum_{s_t} P^\pi(s_0, s_1) \cdots P^\pi(s_{t-1}, s_t) \frac{\partial}{\partial w_i} V_\infty^\pi(s_t)}{N} \ . \tag{5.31}$$

また Fact7より

$$\lim_{N\to\infty} \frac{\sum_{t=0}^{N} \sum_{s_1} \cdots \sum_{s_t} P^\pi(s_0, s_1) \cdots P^\pi(s_{t-1}, s_t) \frac{\partial}{\partial w_i} V_\infty^\pi(s_t)}{N} = \sum_{s} U^\pi(s) \frac{\partial}{\partial w_i} V_\infty^\pi(s) \ . \tag{5.32}$$

よって (5.30)，(5.31)，(5.32) 式より

$$\lim_{N\to\infty} \frac{1}{(1-\gamma)N} \sum_{t=0}^{N-1} \Delta w_i(t) = \sum_{s} U^\pi(s) \frac{\partial}{\partial w_i} V_\infty^\pi(s) \ .$$

■

謝辞

本論文をまとめるにあたり、終始多大なる御指導と御教示をいただきました小林重信教授に心より感謝の意を表します。

本論文に関して適切な御指摘、御意見をいただきました中村清彦教授、伊藤宏司教授、佐藤誠教授に厚く御礼申し上げます。

また、本論文の研究を進める上で多大なる御助言と御教示、御協力をいただきました山村雅幸助教授、宮崎和光助手および小林研究室の皆様に厚く御礼申し上げます。

公表論文

< 学術論文 >

1. 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 (1996)

2. Kimura, H., Miyazaki, K., Kobayashi, S., Policy Improvement by Stochastic Gradient Ascent: A New Approach to Reinforcement Learning in POMDPs, *MACHINE LEARNING* (再投稿予定)

< 国際会議 >

1. Kimura, H., Yamamura, M., Kobayashi, S., Reinforcement learning with delayed rewards on continuous state space, *Proceedings of the 3rd International Conference on Fuzzy Logic, Neural Nets and Soft Computing,* pp.289-292. (1994)

2. 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)

< 国内の学会 >

1. 木村 元，山村 雅幸，小林 重信，状態空間が連続で報酬入力に遅れのある強化学習，計測自動制御学会 第 5 回 自律分散システムシンポジウム (1994)

2. 木村 元，宮崎 和光，小林 重信，確率的傾斜法による強化学習 :不完全知覚問題への接近，計測自動制御学会 システム / 情報合同シンポジウム (1996)

参考文献

[Baird 94] Baird, L. C.: Reinforcement Learning in Continuous Time: Advantage Updating, *Proceedings of IEEE International Conference on Neural Networks*, Vol. IV, pp. 2448-2453 (1994).

[Baird 95a] Baird, L. : Advantage Updating Applied to a Differential Game, *Advances in Neural Information Processing Systems 7 (NIPS-94)*, pp. 353-360 (1994).

[Baird 95b] Baird, L. : Residual Algorithms: Reinforcement Learning with Function Approximation, *Proceedings of the 12th International Conference on Machine Learning*, pp. 30-37 (1995).

[Barnard93] Barnard, E.: Temporal-Difference Methods and Markov Models, *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 23, no.2, March/April 1993, pp. 357-365.

[Barto et al. 83] Barto, A. G., Sutton, R. S. and Anderson, C. W.: Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Problems, *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-13, no.5, September/October 1983, pp. 834-846.

[Barto et.al 93] Barto, A. and Duff, M. : Monte Carlo Matrix Inversion and Reinforcement Learning, *Advances in Neural Information Processing Systems 6 (NIPS-93)*, pp. 687-694 (1993).

[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.

[Boyan et.al 94] Boyan, J. A. and Moore, A. W.: Generalization in Reinforcement Learning: Safely Approximating the Value Function, *Advances in Neural Information Processing Systems 7 (NIPS-94)*, pp. 369-376 (1994).

[Boyan et.al 93] Boyan, J. A. and Littman, M. L.: Packet Routing in Dynamically Changing Networks: A Reinforcement Learning Approach, *Advances in Neural Information Processing Systems 6 (NIPS-93)*, pp. 671-678 (1993).

[Bradtke 93] Bradtke, S. J.: Reinforcement Learning Applied to Linear Quadratic Regulation, *Proc. of NIPS-5* (1993).

[Bradtke et.al 94] Bradtke, S. J. and Duff, M. O.: Reinforcement Learning Method for Continuous-Time Markov Decision Problems, *Advances in Neural Information Processing Systems 7 (NIPS-94)*, pp. 393-400.

[Buckland et.al 93] Buckland, K. M. and Lawrence, P. D.: Transition Point Dynamic Programming, *Advances in Neural Information Processing Systems 6 (NIPS-93)*, pp. 639-646 (1993).

[Cassandra et.al 94] Cassandra, A. R., Kaelbling, L. P. and 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).

[Chapman et.al 91] Chapman, D. and Kaelbling, L. P.: Input Generalization in Delayed Reinforcement Learning: An Algorithm And Performance Comparisons, *Proceedings of the 12th International Joint Conference on Artificial Intelligence*, pp. 726-731 (1991).

[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).

[Crites et.al 95] Crites, R. H. and Barto, A. G.: An Actor/Critic Algorithm that is Equivalent to Q-Learning, *Advances in Neural Information Processing Systems 7 (NIPS-95)*,

pp. 401-408 (1995).

[Dietterich et.al 95] Dietterich, T. G. and Flann, N. S.: Explanation-Based Learning and Reinforcement Learning: A Unified View, *Proceedings of the 12th International Conference on Machine Learning*, pp. 176-184 (1995).

[Dorigo et.al 91] Dorigo, M. and Sirtori, E.: Alecsys: A Parallel Laboratory for Learning Classifier Systems, *Proceedings of the 4th International Conference on Genetic Algorithms*, pp. 296-302 (1991).

[Duff 95] Duff, M. O.: Q-Learning for Bandit Problems, *Proceedings of the 12th International Conference on Machine Learning*, pp. 209-217 (1995).

[Gordon 95] Gordon, G. J.: Stable Function Approximation in Dynamic Programming, *Proceedings of the 12th International Conference on Machine Learning*, pp. 261-268 (1995).

[Heger 94] Heger, M.: Consideration of Risk in Reinforcement Learning, *Proceedings of the 11th International Conference on Machine Learning*, pp. 105-111 (1994).

[Heger 96] Heger, M.: The Loss from Imperfect Value Functions in Expectation-Based and Minimax-Based Tasks, *Machine Learning, 22,*, pp. 197-225 (1996).

[Holland et.al 87] Holland, J. H., and Reightman. J. S.: Cognitive Systems Based on Adaptive Algorithms, Pattern-Directed Inference Systems. Waterman, D. A., and Hayes-Roth, F. ed., Academic Press (1987).

[Holland 86] Holland, J. H.: Escaping brittleness, Machine Learning, an artificial intelligence approach. Volume II. R. S. Michalski, J. G. Carbonell and T. M. Mitchell ed., Morgan Kaufmann, pp. 593-623 (1986).

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

[Jaakkola et.al 93] Jaakkola, T., Jordan, M. I. & Singh, S. P.: On the Convergence of Stochastic Iterative Dynamic Programming Algorithms, *Neural Computation 6*, pp.1185-1201 (1993).

[Kaelbling 93] Kaelbling, L. P.: Hierarchical Learning in Stochastic Domains: Preliminary Results, *Proceedings of the 10th International Conference on Machine Learning*, pp. 167-173 (1993).

[Kimura et.al 94] Kimura, H., Yamamura, M. and Kobayashi, S.: Reinforcement learning with delayed rewards on continuous state space, *Proceedings of the 3rd International Conference on Fuzzy Logic, Neural Nets and Soft Computing* (Iizuka, Japan, August 1-7, 1994) p.p. 289-292.

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

[Liepins et.al 89] Liepins, G. E., Hilliard, M. R., Palmer, M. and Rangarajan, G.: Alternatives for Classifier System Credit Assignment, *Proceedings of the 11th International Joint Conference on Artificial Intelligence*, pp. 756-761 (1989).

[Lin 90] Lin, L. J.: Self-improving Reactive Agents: Case studies of Reinforcement Learning Framework, *Proceedings of the 1st International Conference on Simulation of Adaptive Behavior*, pp. 297-305 (1990).

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

[Lin 93] Lin, L. J.: Scaling Up Reinforcement Learning for Robot Control, *Proceedings of the 10th International Conference on Machine Learning*, pp. 182-189 (1993).

[Littman 92] Littman, M. L.: An optimization-based categorization of reinforcement learning environments, *Proceedings of the 2nd International Conference on Simula-*

tion of Adaptive Behavior (ICSAB), pp. 262-270 (1992).

[Littman 94a] Littman, M. L.: Markov games as a framework for multi-agent reinforcement learning, *Proceedings of the 11th International Conference on Machine Learning*, pp. 157-163 (1994).

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

[Littman et.al 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).

[Mahadevan 92] Mahadevan, S.: Enhancing Transfer in Reinforcement Learning by Building Stochastic Models of Robot Actions, *Proceedings of the 9th International Conference on Machine Learning*, pp. 290-299 (1992).

[Mahadevan 94] Mahadevan, S.: To Discount or not to Discount in Reinforcement Learning: A Case Study Comparing R Learning and Q Learning, *Proceedings of the 11th International Conference on Machine Learning*, pp. 164-172 (1994).

[McCallum 92] McCallum, R. A.: Using Transitional Proximity for Faster Reinforcement Learning, *Proceedings of the 9th International Conference on Machine Learning*, pp. 316-321 (1992).

[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 95a] McCallum, R. A.: Instance-Based State Identification for Reinforcement Learning, *Advances in Neural Information Processing Systems 7 (NIPS-95)*, pp. 377-384 (1995).

[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).

[Miyazaki et.al 94] Miyazaki, K., Yamamura, M. and Kobayashi, S.: On the Rationality of Profit Sharing in Reinforcement Learning, *Proceedings of the 3rd International Conference on Fuzzy Logic, Neural Nets and Soft Computing* (Iizuka, Japan, August 1-7, 1994) p.p. 285-288.

[Moore 94] Moore A. W.: Prioritized Sweeping: Reinforcement Learning With Less Data and Less Time, *Machine Learning 13*, pp. 103-129 (1994).

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

[Peng et.al 92] Peng, J. and Williams, R. J.: Efficient Learning and Planning Within the Dyna Framework, *Proceedings of the 2nd International Conference on Simulation of Adaptive Behavior (ICSAB)*, pp. 281-290 (1992).

[Peng et.al 94] Peng, J. and Williams, R. J.: Incremental Multi-Step Q-Learning, *Proceedings of the 11th International Conference on Machine Learning*, pp. 226-232 (1994).

[Peng 95] Peng, J. : Efficient Memory-Based Dynamic Programming, *Proceedings of the 12th International Conference on Machine Learning*, pp. 438-446 (1995).

[Schwartz, A. 93] Schwartz, A.: A Reinforcement Learning Method for Maximizing Undiscounted Rewards, *Proceedings of the 10th International Conference on Machine Learning*, pp. 298-305 (1993).

[Sen et.al 94] Sen, S., Sekaran, M. and Hale, J.: Learning to coodinate without sharing information, *Proceedings of the 12th National Conference on Artificial Intelligence*, Vol. 1, pp. 426-431 (1994).

[Singh 92] Singh, S. P.: Transfer of Learning by Composing Solutions of Elemental Sequential Tasks, *Machine Learning 8*, pp. 323-339 (1992).

[Singh 94] Singh, S. P.: Reinforcement Learning Algorithms for Average-Payoff Markovian Decision Processes, *Proceedings of the 12th National Conference on Artificial Intelligence*, Vol. 1, pp. 700-705 (1994).

[Singh et.al 94] Singh, S. P., Jaakkola, T. and 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).

[Singh et.al 94] Singh, S. P., Jaakkola, T. and Jordan, M. I.: Reinforcement Learning with Soft State Aggregation, *Advances in Neural Information Processing Systems 7 (NIPS-94)*, pp. 361-368.

[Singh et.al 96] Singh, S. P. and Sutton, R.S.: Reinforcement Learning with Replacing Eligibility Traces, *Machine Learning 22*, pp. 123-158 (1996).

[Sutton 88] Sutton, R. S.: Learning to Predict by the Methods of Temporal Differences, *Machine Learning 3*, pp. 9-44 (1988).

[Sutton 90] Sutton, R. S.: Reinforcement Learning Architectures for Animats, *Proceedings of the 1st International Conference on Simulation of Adaptive Behavior*, pp. 288-295 (1990).

[Sutton 95] Sutton, R. S.: TD Models: Modeling the world at a Mixture of Time Scales, *Proceedings of the 12th International Conference on Machine Learning*, pp. 531-539 (1995).

[Tan 91] Tan, M.: Cost-Sensitive Reinforcement Learning for Adaptive Classification and Control, *Proceedings of the 9th National Conference on Artificial Intelligence (AAAI91)*, pp. 774-780 (1991).

[Tan 93] Tan, M.: Multi-Agent Reinforcement Learning: Independent vs. Cooperative Agents, *Proceedings of the 10th International Conference on Machine Learning*, pp. 330-337 (1993).

[Tesauro 92] Tesauro, G.: Temporal Difference Learning of Backgammon Strategy, *Proceedings of the 9th International Workshop on Machine Learning*, pp. 451-457 (1992).

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

[Whitehead et.al 90] Whitehead, S. D. & Ballard, D. H.: Active Perception and Reinforcement Learning, *Proceedings of the 7th International Conference on Machine Learning*, pp. 179-188 (1990).

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

[Williams 87] Williams, R. J.: A Class of Gradient-Estimating Algorithms for Reinforcement learning in Neural Networks, *IEEE First International Conference on Neural Networks*, volume II, pp. 601-608 (1987).

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

[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).

[宮崎 94] 宮崎 和光，山村 雅幸，小林 重信: 強化学習における報酬割当の理論的考察, 人工知能学会誌, Vol.9, No.4, pp.580-587 (1994).