POMDPs 環境のための報酬獲得効率に基づく強化学習法

A Reinforcement Learning Method Using Reward Acquisition Efficiency for POMDP Environments

河合 宏和
Hirokazu Kawai

大阪市立大学大学院工学研究科電子情報系専攻 *1
Department of Physical Electronics and Informatics, Graduate School of Engineering, Osaka City University
hiro@kdel.info.eng.osaka-cu.ac.jp

上野 敦志
Atsushi Ueno

(同 上)
ueno@info.eng.osaka-cu.ac.jp

辰巳 昭治
Shoji Tatsumi

(同 上)
tatsumi@info.eng.osaka-cu.ac.jp

keywords: reinforcement learning, POMDPs, reward acquisition efficiency, multi-reward environment

Summary

Reinforcement Learning (RL) methods are very hopeful because they can learn useful behavior based on rewards from environment by trial and error. This paper tackles more difficult problems than the ones tackled by many ordinary RL methods: RL in POMDP (Partially Observable Markov Decision Process) environments with multiple rewards.

Environments in which RL is performed are usually modeled as MDP (Markov Decision Process) environments. A POMDP environment is an extension of an MDP environment, which is not always fully observable by agents in it. Learning in a POMDP environment tends to be very difficult because the agent may perceive the same observation for some different environmental states.

Many ordinary RL methods are designed to be performed in an environment with a single reward, which gives a definite amount of reward at a single definite state. However, in many real problems, the agent must select among multiple destination states each of which gives a certain amount of reward different from each other. In such case, the agent must evaluate them by Reward Acquisition Efficiency (RAE).

In this paper, we propose a new RL method for POMDP environments with multiple rewards. First, we introduce a condition proposed by Uemura which is useful for RL methods to learn effectively in POMDP environments. Next, we draw attention to an RL method applicable to multi-reward environments, RAE-RL, which uses the averaged RAEs as the values of rules. And we show that RL methods which use the average amounts as the values, like RAE-RL, are effective in POMDP environments.

Then, we extend the condition for POMDP environments. Our proposal method is developed by applying it to an RL method using RAE. Finally, we have verified that our proposed method can learn effectively in POMDP environments with multiple rewards by computer simulations.

1. は じ め に

強化学習 (RL: Reinforcement Learning) は、試行錯誤により行われる教師なし学習の一種である。学習者はエージェントと呼ばれる。各時刻において、エージェントはセンサなどによって環境の状態 s ($s \in S$: S は状態の集合) を観測 o ($o \in O$: O は観測の集合) として知覚し、現在の観測において実行可能な行動群の中から 1 つを選んで実行する。この観測と行動の対をルールと呼び、観測から行動への変換関数を政策と呼ぶ。観測 o で

行動 a を選択するルールを (o, a) と書く。各ルールは価値と呼ばれる重みを持ち、報酬情報を基に価値を更新することで学習は進む。

強化学習の目的は、単位時間あたりの報酬獲得量を最大化することにある*2。無限に続くタスク (連続タスク) の場合には、報酬獲得量の合計が発散するので、報酬を単位時間ごとに割り引いたものの合計 (割引報酬と呼ばれる [木村 96]) の期待値を計算し、その最大化を目的

*1 現在、株式会社ダイヘン勤務

*2 これは、一種の時給とみなすことができる。他にも単位移動距離あたり、単位消費エネルギーあたりなどを考えた方が適切な場合もあるが、簡単のために、本論文ではそれらの変化率が一定であるとして同一視する。

とすることが一般的である。しかし、報酬獲得量の合計が有限であるタスクの場合は、報酬の時間平均の期待値（平均報酬と呼ばれる [木村 96]）を最大化することが適切である [Mahadevan 96]。このようなタスクの代表としては、エージェントがゴール状態に到達した時に報酬が得られ、スタート状態（の 1 つ）に戻されるようなタスク（エピソード的タスク）がある。多くの強化学習法では割引報酬の期待値の最大化を目的としているが、エピソード的タスクに適用すると、問題が発生する場合がある*3。この問題が顕著に現れる例として、複数のゴール状態でそれぞれ異なる量の報酬が与えられる環境が考えられる（[Mahadevan 96] 参照）。このような環境を複数報酬環境と呼ぶことにする。また、強化学習で扱う問題環境のうち、エージェントの知覚能力に限界がある問題環境を部分観測マルコフ決定過程（POMDPs: Partially Observable Markov Decision Processes）環境といい、POMDPs 環境で起こる問題を不完全知覚問題という。

本論文の目的は、複数報酬を持つ POMDPs 環境のための新しい強化学習法の提案である。そのために、まずそれぞれの問題に対応する従来研究を紹介する。POMDPs 環境に特化した強化学習法としては、EPS (Episode-based Profit Sharing) [植村 05] が提案されている。複数報酬環境に対しては、報酬獲得効率 (RAE: Reward Acquisition Efficiency) に基づく手法が提案されている [内田 99, 山口 00, 石村 01]。本論文では、[植村 05] で示されている POMDPs 環境に適応するための条件を拡張し、報酬獲得効率に基づく手法に適用して開発した新たな強化学習法を提案する。そして両方の利点を組み合わせることで、提案手法が複数報酬を持つ POMDPs 環境において従来手法よりも有効であることを示す。

以下、2 章では不完全知覚問題について説明し、EPS を紹介する。3 章では複数報酬環境における一般的な強化学習法の問題点を挙げ、そのような環境で適切な学習を行える RAE-RL [内田 99] について説明する。4 章では各ルールへの報酬の配量（以下、強化値）の平均をルールの価値として用いる学習法の性質を分析する。5 章では報酬獲得効率に基づいた新たな学習法を提案し、不完全知覚問題に対する頑健性が向上することを実験により確かめる。

2. 不完全知覚問題

強化学習では、エージェントはセンサなどで環境の状態 s を観測 o として知覚する。MDPs (Markov Decision Processes) 環境のように環境を完全に知覚できる場合は $s = o$ となるが、エージェントの知覚能力に制限がある POMDPs 環境では、複数の異なる状態を同一の観測と

見なすことで、状態が混同される可能性がある。このように、異なる状態が混同される問題は不完全知覚問題の一種である。[宮崎 99] および [Arai 01] では、不完全知覚によって生じる問題の分類が提案されている。

2.1 ブートストラップ型の学習法において生じる問題

あるルール（または観測）の価値を計算する際に、他のルール（または観測）の価値を基準として使用する*4 学習法を、ブートストラップ型と呼ぶ。この場合、状態の混同の影響により、ルール（または観測）の価値が振動して収束しなくなる可能性がある。また、この振動が他のルール（または観測）の価値に伝播し、学習全体が適切に収束しないおそれがある。ブートストラップ型の学習法としては、Sarsa(λ) が不完全知覚問題に頑強であるといわれている [Loch 98]。しかし、混同された観測で複数の行動を選択する必要が生じる（すなわち、最適なルールが分散している）ような問題（2.2 節参照）では、Sarsa(λ) でも適切に学習することができないことが経験的に示されている [Arai 01]。このようなタイプの問題は、不完全知覚問題の中でも特に困難である。本論文では、このタイプの問題を最適ルール分散型の問題と呼び*5、以降、このタイプの不完全知覚問題を扱う。

最適ルール分散型の問題に対して、観測の遷移の履歴を用いることによって状態を見分けるという対策を取ることができる。この種の方法をメモリーベース法と呼ぶ。適切な長さの履歴を用いることによってうまく状態を分離することができれば、問題を MDPs として扱うことができる。しかし、適切な履歴の長さをあらかじめ決定するのが困難であること、また履歴に含まれる観測の順列の数が指数的に発散し、時間的および空間的な計算量が発散することなどの問題がある（[宮崎 03] 参照）。従って、本論文では遷移の履歴を用いない方法のみを扱う。

Profit Sharing 法（以下、PS）[Grefenstette 88, 宮崎 94] のような非ブートストラップ型の学習法は価値計算に他の観測、ルールの価値を使わないため、不完全知覚問題に対し比較的頑健であると言われている。PS を改良した FVPS (FirstVisit PS) [Arai 01] や EPS は、最適ルール分散型の問題において、Sarsa(λ) よりもかなり良い性能を示している。従って、本論文ではこれ以降、非ブートストラップ型の学習法のみを扱う。

実際のところ、PS も最適ルール分散型の問題にはうまく対応できない。次節では、この問題とその対策について説明する。

*3 一方、POMDPs 環境（後述）で表される連続タスクの場合には、割引報酬の期待値の最大化が平均報酬の最大化に一致することが示されている [Singh 94]。

*4 主に、遷移先の状態でのルール（または観測）の価値を基準として使用する。

*5 [Arai 01] において「confusion type (D)」と呼ばれている最も難しいカテゴリの混同問題と同一である。

図 1 見かけ上の無効ルールが発生する環境

2.2 最適ルール分散型の問題

PS では、一般的に強化値の累積値を価値として用いる。従って、PS は累積値ベース [高玉 03] の強化学習法の一つである。PS において、常にループ系列^{*6}上にあるルールである無効ルールを抑制する条件として、合理性定理が提案されている [宮崎 94]。この合理性定理に従ってルールの強化を行う際、状態の混同によって実際は無効ルールでないルールを無効ルールと見なし、報酬獲得に必要なルールが抑制される場合がある。

例として図 1 に示す環境について考える。この環境では、状態 s_1 と s_2 が混同され、共に観測 o として知覚される。すなわちこの環境は実際には図 1(a) に示す環境であるにも関わらず、エージェントには図 1(b) のように知覚される。報酬を獲得するためには、状態 s_1 では行動 b を、状態 s_2 では行動 a を取る必要がある。この際、合理性定理に従った学習を行うと、エージェントが知覚する環境において無効ルールと見なされるルール (o, b) は抑制対象となり、状態 s_1 においてループ系列を形成しているルール (o, a) のみが強化される。

このような問題に対して、確率的な行動選択を取り複数の行動を選択する政策を得ることで対応する手法が考えられ、確率的傾斜法 [木村 96]、FVPS、PS-r^{*} [宮崎 03]、EOPSwJ (Extended On-line PS with Judgement) [斎藤 04]、EPS などが提案されている。このうち PS-r^{*} は、複数の行動を選択する必要がある観測において全行動からランダム選択をするため、必要な行動のみからランダム選択をおこなう EPS の方が有効である [植村 05]。EOPSwJ も同様に、全行動からのランダム選択という弱点を持つ。確率的傾斜法は、割引報酬の最大化に特化した手法であるため、複数報酬環境向けに改良することが困難である。FVPS と EPS は非常に良く似た手法で、FVPS の強化関数を重ね合わせたものが EPS の強化関数であるとみなすことができる^{*7}。次節では、本論文で議論の対象としている EPS を紹介し、EPS がこの問題に対しどのような解決を行っているかを説明する。

2.3 Episode-based Profit Sharing

EPS では、エージェントが報酬を獲得した際、それまでの行動系列が一括して強化される。この行動系列をエピソードと呼ぶ。報酬から x 番目のルール (o_x, a_x) における価値の更新は以下の式によって行われる。ただし、エピソード長を W とし、 $x = 1, 2, \dots, W$ とする。

$$Q(o_x, a_x) \leftarrow Q(o_x, a_x) + R \times f(x) \quad (1)$$

ここで、 $Q(o_x, a_x)$ はルール (o_x, a_x) の価値、 R は獲得された報酬値、 $f(x)$ は報酬から x 番目のルールへの報酬分配法を表す強化関数である。一般に、強化関数は常に正の値をとる。記号は [植村 05] とは変えてあり、以降、他の手法でも同じ記号を用いる。この式は、通常の PS と同じであり、累積値ベースの学習法における価値の更新式の一般形といえる。

[植村 05] では、POMDPs 環境中で、「不完全知覚問題の影響により、報酬獲得に複数のルールの選択を必要とする観測において、ランダム選択に劣らない性能を持つためには、報酬獲得に必要なルールを、すべて同じ確率で選択できれば十分である」ことが示されている。すなわち、決定論的な状態遷移を持つ POMDPs 環境では、状態の混同が起こる観測で報酬獲得に必要なルールを均一に強化すれば、最適ルール分散型の問題に対応することができる。[植村 05] では、この問題に対して、「1 エピソード中の同一観測上のルールに対して、報酬を等しく分配すること」という条件を満たすことで対応している^{*8}。具体的には、次式の強化関数を用いている。

$$f(x) = \begin{cases} \alpha_{o_x} & \text{このルールの強化が初めての場合} \\ 0 & \text{それ以外の場合} \end{cases} \quad (2)$$

ここで、 α_{o_x} は各観測 o_x ごとに一定である。本論文では、この条件を均一分配と呼ぶことにする。累積値ベースの学習法では、後は強化値を累積するだけなので、この条件を満たせば、状態の混同が起こる観測で報酬獲得に必要なルールが均一に強化される。

ただしこの条件は、スタート地点を固定にした場合に、均一に強化することを保証するものである。エピソードごとに異なる状態からスタートする問題の場合には、報酬獲得に必要なルールがエピソードごとに变化する。従って、全てのスタート地点から報酬獲得に必要なルールを、各観測ごとに均一に強化することは、非常に困難である。以後、本論文では全エピソードのスタート地点が同一である問題を議論の対象とする。

3. 複数報酬環境における問題

強化学習の目的とは、設計者によって設定された報酬の単位時間あたりの獲得量の期待値（平均報酬）を最大

^{*6} 同じ状態に戻り、報酬獲得に不要な行動系列のこと。[宮崎 94] では、う回系列と呼んでいる。

^{*7} [植村 05] 中の single-EPS と呼ばれる手法が、FVPS とほとんど同じである。

^{*8} FVPS でも同様の対策を取っている。

化することである．Q-learning [Watkins 92], PS など、多くの学習法では、この平均報酬の最大化という目的を割引報酬の期待値の最大化で近似している．割引報酬に基づいた学習法では、単位時間の経過ごとに報酬を割引率 γ ($0 < \gamma < 1$) で割り引いた値を各ルール（または観測）に分配し、価値計算を行う．しかし 1 章で述べたように、エピソード的タスクの場合には、割引報酬による近似が適切ではなく、問題が生じる場合がある．エピソード的タスクは、迷路問題のみではなく、ゲームやロボット制御など様々な分野で強化学習問題として扱われており [Mahadevan 96]、この問題を無視することはできない．

このような問題が顕著に現れる例としては、複数報酬環境がある*9．複数報酬環境下で単位時間あたりの報酬獲得量を最大化するためには、報酬への距離（ステップ数）と報酬の大きさのバランスを考慮して、最も効率のよい報酬への経路を学習する必要がある．しかし、割引報酬に基づいた学習法では、割引率 γ が 1 に近づきすぎると、学習の目的が報酬の大きさに偏りすぎて報酬への距離を軽視した政策が学習される．逆に、割引率 γ が 0 に近づきすぎると、学習の目的が報酬までの距離に偏りすぎて報酬の大きさを軽視した政策が学習される．このような問題は、適切な割引率の値設定が問題環境に依存するため、解決が非常に困難である．具体的には、適切な割引率は各報酬までの距離および経路上の遷移確率に依存して決まるので、学習を行う前に決定することはできず、値を変えて何度も試行錯誤するしかない．また PS の場合では、ループ系列を抑制するために合理性定理を満たすよう γ を制限すると、大きな（1 に近い）値には設定できないので、値設定がますます困難（問題によっては不可能）になる．

このような問題を回避する学習法である RAE-RL や RAE-PIA [山口 00, 石村 01] では、各ルールの価値を報酬獲得効率で表現する．報酬獲得効率は、移動距離（および時間）に反比例した強化値を平均することによって求められる．すなわち、そのルールを実行後の平均報酬に相当し、価値の高いルールを選択することで、平均報酬の最大化が期待できる．

本論文では、非ブートストラップ型である RAE-RL を議論の対象とする．

3.1 RAE-RL

RAE-RL では、エピソード単位での学習を行う．報酬を獲得するとエピソード中の行動系列が一括で強化される．強化関数 $f(x)$ は、以下に示すように、報酬までの距離 x に反比例する*10．

$$f(x) = \frac{1}{x} \quad (3)$$

*9 単一報酬でも状態遷移が確率論的な環境では、同様の問題が起り得る．

*10 記号は [内田 99] とは変えてある．式(4) も同様．

表 1 非ブートストラップ型の学習法の分類

価値の 計算法	報酬の 分配法	
	割引型	反比例型
累積値ベース	PS EPS	
平均値ベース		RAE-RL

累積値ベースの PS では、報酬獲得ごとにエピソード中の行動系列に強化値が累積されていく．それに対し RAE-RL は、LMS 法 (Least Mean Squares) [Russell 95] をベースとし、式(3) の強化関数に従って分配された強化値の平均（以下、平均強化値と呼ぶ）を算出し各ルールの価値とする．

$$\begin{aligned} I(o_x, a_x) &\leftarrow I(o_x, a_x) + 1 \\ Q(o_x, a_x) &\leftarrow \left(1 - \frac{1}{I(o_x, a_x)}\right) Q(o_x, a_x) \\ &\quad + \frac{R \times f(x)}{I(o_x, a_x)} \end{aligned} \quad (4)$$

ここで、 $I(o_x, a_x)$ はルール (o_x, a_x) の強化回数である．式(4) は平均値ベースの学習法における価値更新式の一般形といえる*11．

RAE-RL では、報酬の分配法としては距離（および時間）に反比例し（反比例型）、分配報酬からの価値の計算法としては、平均を取る（平均値ベース*12）．通常の PS が単位距離（および時間）ごとに報酬を割り引いて分配し（割引型）、その分配報酬を累積する（累積値ベース）のと比較すると、表 1 のようにまとめることができる．前述のように、反比例型の報酬分配は複数報酬環境への適応のために非常に重要である．一方、平均値ベースの価値計算も、累積値ベースのものと比較すると、いくつかの性質の変化が見られる．特に重要なのは、最適ルール分散型の問題への頑健性の変化である．4 章では平均値ベースの学習法が持つ性質について、この変化を含めて詳しく分析する．

4. 平均値ベースの学習法が持つ性質

累積値ベースの学習法では、強化値をルールごとに単純加算するため、正の報酬のみを扱う環境や負の報酬のみを扱う環境で、 $f(x) > 0$ の条件を満たす時、各ルールの価値は学習の進行と共にそれぞれ発散していくことになる．一方、平均値ベースの学習法では、 $f(x)$ が有限の値を取る場合、ルールの価値が発散することは無い．

また、正の報酬のみを扱う場合、累積値ベースの学習法で $f(x) > 0$ を満たす時、報酬獲得時にエピソード中の

*11 例えば、 $f(x) = 1$ とすれば、LMS 法となる．

*12 これは、[高玉 03] において、推定値ベースと呼ばれているものの一種であると考えられる．推定値ベースでは、忘却を考えて、重み付きの平均を取る．

図 2 無効ルールの抑制が最も困難な状態

各ルールの価値は必ず増加する．それに対し平均値ベースの学習法では獲得された強化値を平均化するため，正の報酬のみを扱う場合でも報酬獲得時にエピソード中の各ルールの価値が減少することもある．

以下に，累積値ベースの学習法と平均値ベースの学習法について，ループ系列の強化と最適ルール分散型の問題で生じる差異を検討する．

4.1 ループ系列の強化に関する性質

累積値ベースの PS や EPS では，正の報酬のみを持つ環境において，獲得された分配報酬がルールの価値に単純加算されるため，強化回数の多いルールほど高く評価されやすいという傾向がある．そのため，報酬獲得に貢献しない無効ルールが集中的に選択されやすいような環境においては，無効ルールを強く抑制し，無効ルールでないルール（有効ルール）を強化するような強化関数を使わなくては報酬獲得が阻害される可能性がある．[宮崎 94] では，行動選択数を L とした時に，無効ルールの抑制が最も困難な状態として，図 2 のような

$L-1$ 本の有効ルールと 1 本の無効ルールが

混在する状態

が示されている．このような環境では，無効ルールが集中的に強化される一方で有効ルールが分散的に強化されることで，強化回数の差によって無効ルールが有効ルールよりも高く評価される可能性がある．このような環境でも無効ルールを常に抑制するためには，無効ルール抑制条件である合理性定理を満たす必要がある．

しかし，平均強化値を用いた場合，強化値の合計を強化回数で除算するため，このような強化回数の偏りによる影響を受けず，無効ルールも含めて各ルールをより適切に評価することが期待できる．

一方，学習途中においては各ルールの強化回数も少なく，平均強化値が不安定であることが多いため，ループ系列が一時的に強化される恐れがある．エピソードごとに更新を行う学習法ではループ系列を一時的にでも強化すると学習が進まなくなる可能性もある．その点を考慮し，学習初期の行動選択にはランダム性を強くする必要がある．また，学習が進み強化回数が増えれば適切な価値決定が期待できるが，合理性定理のように無効ルールの確実な抑制が保証されるわけではない．

図 3 Sutton の迷路問題 [Sutton 90]

4.2 最適ルール分散型の問題に関する性質

累積値ベースの学習法では，正の報酬のみを扱う環境において，ルールの価値が下がるということはない．そのため合理性定理に基づく無効ルールの抑制は，無効ルールの価値を直接下げるという形では行われず，ループ系列から脱出する有効ルールを強く強化するという形で実現されている．そのため，図 1 に示すような，不完全知覚問題により見かけ上の有効ルール (o, a) が実際は無効ルールも含んでいるような環境では，見かけ上の無効ルール (o, b) を抑制するために見かけ上の有効ルールを強化し，その結果ループ系列に陥いる可能性がある．

一方，平均強化値を用いた場合は，ルールの価値が下がることもある．一般に強化関数には非増加関数を用いるため，以下の関係が成り立つ．

$$f(x_i) \geq f(x_k) \quad (x_i < x_k) \quad (5)$$

これにより，無効ルールが同一エピソード中でくり返し選択されると，回数に応じて無効ルールの得る強化値の平均は下がっていくと考えられる（ただし，LMS 法の場合は，式(5)で常に等号が成り立つので，このような現象は起こらない）．そのため，仮に図 1 のような環境で，見かけ上の有効ルールが強化されたとしても，その見かけ上の有効ルールが連続的に選択されるとその価値は下がっていくと考えられる．結果として，最適ルール分散型の問題の影響により複数行動の選択が必要な観測において，累積値ベースの学習法よりも高い頑健性が期待できる．

4.3 実験

平均値ベースの学習法が持つ性質を，累積値ベースの学習法との比較実験で確認する．

§1 ループ系列の強化に関する実験

平均値ベースの学習法のループ系列の強化に関する性質を，図 3 の迷路走行タスク [Sutton 90] で確認する．

まず，問題環境の設定を説明する．それぞれのマスの状態とし，エージェントは観測として自身の居るマスを確認する．上下左右に隣接する 4 マスへの移動を行動とし，黒いマスで表される壁や環境の外側への移動を選択

すると移動は行われない．状態 S から学習を開始し，状態 G に到達すると 10 の報酬を得て状態 S へと戻る．エピソード単位で学習を行う手法では，この際に状態 S から G へと至るまでの行動系列が強化される．単位行動あたりの報酬獲得量を性能の評価とする．乱数を変えた実験を 100 回行い，性能の平均値を求める．

報酬までの最短経路は 14 ステップであるため，最適解は $10/14 \simeq 0.714$ である．このような実験環境で，強化関数を反比例型とした PS (以下，反比例型の PS と呼ぶ) と RAE-RL とを比較する．反比例型の PS では，報酬獲得時に報酬から x 番目のルール (o_x, a_x) の価値 $Q(o_x, a_x)$ を以下の更新式に従って更新する．

$$f(x) = \frac{1}{x}$$

$$Q(o_x, a_x) \leftarrow Q(o_x, a_x) + R \times f(x) \quad (6)$$

この $f(x)$ は，[宮崎 94] で示された合理性定理の条件を満たしていないため，ループ系列を抑制できない場合がある．一方，RAE-RL では，強化関数は全く同じで，価値の更新で平均を取るという点だけが異なっている (式(4)参照)．この比較は，累積値ベースの価値更新を平均値ベースに変えるだけで，ループ系列の抑制効果が高まることを示すのが目的である．

両手法において，行動選択法はボルツマン選択を用いる．PS では，通常，ルーレット選択が良く用いられるが，ルーレット選択ではルールの価値の比のみに依存して選択確率が決まるので，最適ルールに近い価値を持つルールが存在すると，ルール選択がいつまでも収束しない．RAE-RL のように報酬獲得効率で価値を表現する場合 (特に報酬から遠い所では) 価値の比があまり大きくならないので，ルール選択が収束しないことになる^{*13}．ボルツマン選択の場合には，温度パラメータ T を調整することで，ルール選択のランダム性を調整することができる．一般的には，学習の進行と共に T の値を下げていくことで，ランダム性を低くしていく．この実験でもこれを採用する．また，設計の簡単のため，観測ごとにルールの価値を正規化して和が 1 になるような値をボルツマン選択の重みとして用いる．これにより選択確率が価値の比率と温度パラメータ T のみに依存するようになる (以下，このような選択法を正規ボルツマン選択と呼ぶ)．具体的には， T は以下の式に従って下げていく．

$$T = \max(T_s \times \Gamma^t, T_m) \quad (7)$$

ここで， t は学習開始時からの総ステップ数を表し， T_s は T の初期値， Γ は T の割引率， T_m は T の最小値を表すパラメータである．以降の実験においても，正規ボルツマン選択の温度パラメータ T には，この式を用いる．この実験においては，いくつかの値を試してみて，両手法とも $T_s = 1.0$ ， $\Gamma = 0.9997$ ， $T_m = 0.01$ とした．

*13 累積値ベースの価値計算の場合には，同じルールを選択し続けると価値が発散するので，そのルールに収束する．

図 4 Sutton の迷路問題における実験結果

図 5 最適ルール分散型の問題を含む迷路問題

この実験の結果を図 4 に示す．強化関数 $f(x) = 1/x$ は合理性定理を満たしていないため，累積値ベースの学習法に適用した場合，ループ系列を強化し，報酬獲得が阻害されることがあった．それに対し RAE-RL では，同じ強化関数を用いているにもかかわらずループ系列に対して適切に学習が行われているため，結果として学習性能に差が生じた．このことは，価値として平均値を用いることが，ループ系列の抑制に有効であることを示している．

§2 最適ルール分散型の問題に関する実験

ここでは，平均値ベースの学習法が持つ最適ルール分散型の問題に関する性質を図 5 に示す問題環境で確かめる．問題環境の設定は，前項の実験で述べたものを使用し，状態 G に到達した時の報酬量は 10 とする．それに加え，この問題環境では不完全知覚問題により，同一観測として混同される異なる状態が存在する． A と書かれた 4 状態が不完全知覚の影響を受け，同一の観測として見なされる．この観測 A では，下と右という複数の行動を選択する必要がある．この問題環境において，観測 A では報酬獲得に必要な行動である下と右を等確率で選択し，その他の観測では最適行動を取らせる政策が最適である．この政策を実行させたところ，約 0.911 という性能値が得られた．

このような環境で，最適ルール分散型の問題に対応した手法である EPS，合理性定理を満たした累積値ベースの PS，PS における各ルールの価値を平均値とした手法 (以下，平均値ベースの PS と呼ぶ) の比較実験を行った．

図 6 最適ルール分散型の問題を含む迷路問題における実験結果

表 2 観測 A における行動選択確率

行動\手法	累積値ベースの PS	平均値ベースの PS	EPS
上	0.053	0.072	0.117
下	0.032	0.121	0.382
左	0.003	0.007	0.119
右	0.912	0.800	0.382

簡単のため、行動選択法はルーレット選択を使用する。累積値ベース、平均値ベースの PS における強化関数は $f(x) = (1/4)^x$ とした。行動選択数が 4 であるため、この強化関数は [宮崎 94] の合理性定理を満たしている。

この実験の結果を図 6 に示す。累積値ベースの PS では、ランダム選択の性能とほぼ同等の結果となった。それに対し PS を平均化することで性能の向上が見られた。この問題環境において、実験終了時の観測 A における各手法の行動選択確率を表 2 に示す。累積値ベースの PS では、状態の混同が起こる観測 A において、約 0.91 の確率で右行動を選択し、報酬獲得に必要なもう 1 つの行動である下行動の選択確率は約 0.03 にまで落ち込んでいる。このため、観測 A のうちの下行動が必要な 2 状態において、ループ系列から中々脱出できず性能の低下を招いたと考えられる。それに対し、平均値ベースの PS では、観測 A における右行動の選択確率は約 0.80 に下がり、下行動の選択確率は約 0.12 に上がっている。このため、観測 A においてループ系列に陥りにくくなることで、累積値ベースの PS に対して性能が上昇したと考えられる。この実験結果から、価値として平均値を用いることが、最適ルール分散型の問題に対して有効であることが分かる。ただし、EPS では観測 A において報酬獲得に必要な行動である右行動と下行動が均一に強化され、平均値ベースの PS 以上の最適ルール分散型の問題への頑健性が確認できる。

5. POMDPs 環境のための報酬獲得効率に基づく強化学習法

前章で示したように、平均値ベースの学習法は最適ルール分散型の問題に対してある程度の頑健性を持っていることが分かった。これを [植村 05] で示された最適ルール分散型の問題に適応するための条件を満たすように変換することで、さらなる頑健性を得ることが期待できる。また、報酬獲得効率に基づくことで、複数報酬環境に対応することが期待できる。本章では、この観点から、報酬獲得効率に基づく学習法を POMDPs 環境に拡張した新たな学習法を提案する。

5.1 提案手法

§1 アルゴリズム

POMDPs 環境中で累積値ベースの学習法の強化関数が持つべき条件として、均一分配という条件が示されている [植村 05]。この条件を一般の非ブートストラップ型の学習法に適用可能なように拡張するためには、分配後の価値の計算部分も等しくする必要がある。従って、一般の非ブートストラップ型の学習法が満たすべき条件は、「1 エピソード中の同一観測上のルールに対して、価値を等しく更新すること」と表すことができる。以降、この条件を均一更新と呼ぶことにする。

まず簡単のために、状態の混同が起こる観測で、報酬を得るために必要なルールの集合が変化しない場合を考える（次項では、この制限を緩めた場合を考察する）。このような環境では、各試行で報酬を得るためには、集合内のルールが 1 度以上選択される必要がある。

このような場合に均一更新を満たすためには、エピソード中の同一観測上の各ルールに対して、等しい強化値を一度だけ分配し、価値を一度だけ更新すればよい。累積値ベースの学習法では、強化値 0 で更新を行っても価値が変化しないので、均一分配と均一更新は同じことを意味する^{*14}。均一更新を行う手法としては、古典的な初回訪問モンテカルロ法 [Singh 96] がある。ただし、この手法は報酬までの距離を全く考慮しないという点で、一般的な強化学習法とは異なる。

均一更新の条件を報酬獲得効率に基づく学習法に適用する。平均値ベースの学習法では、ルールの価値は (強化値の合計/強化回数) の形で得ることができる。均一分配の条件を満たした強化関数では、報酬獲得に必要なルールの強化値の合計は観測ごとに均一になる。これに加え、強化回数を均一にすることで、均一更新の条件を満たし、報酬獲得に必要なルールの価値を均一にすることができる。

POMDPs 環境のための報酬獲得効率に基づく学習法

$x = W$ から開始し、 $x = 1$ まで順に以下の操作を行う。

*14 均一分配を行わない学習法でも、エピソード中で複数回出現したルールの価値更新を一度しか行わない手法が存在する [松井 03]

- (1) 観測 o がエピソード中に初めて現れた時、その時の x を x_o として、観測 o 上のルールのための報酬分配 $F(o)$ を以下のように決定する。

$$F(o) = \frac{1}{x_o} \quad (8)$$

- (2) ルール (o, a) がエピソード中に初めて現れた時、ルールの価値 $Q(o, a)$ 、強化回数 $I(o, a)$ を次式で更新する。

$$\begin{aligned} I(o, a) &\leftarrow I(o, a) + 1 \\ Q(o, a) &\leftarrow \left(1 - \frac{1}{I(o, a)}\right) Q(o, a) \\ &\quad + \frac{R \times F(o)}{I(o, a)} \end{aligned} \quad (9)$$

- (3) 2 度目以降にルールが現れた時は、更新を行わない。従って、エピソード中の各ルールは、それぞれ 1 度だけ更新されることになる。

式(8) は RAE-RL と同一の反比例型の強化関数であり^{*15}、式(9) は式(4) と同一の平均値ベースの価値更新式である。従って、この提案手法における各ルールの価値は報酬獲得効率であり、複数報酬環境に適応可能である。

ただし、 $F(o)$ は、RAE-RL の (および通常の) 強化関数 $f(x)$ とは異なり、観測 o の関数であり、エピソード中の同じ観測上のルールは全て等しい強化値を受け取る。また、ステップ (3) で示したように、強化回数も等しくなる。従って、均一更新の条件を満たし、POMDPs 環境への適応性が期待できる。

$1/x$ は x に対する減少関数であるため、各ルールの強化値 $R \times F(o)$ は、RAE-RL で学習した場合の強化値 $R \times f(x)$ と同じかそれ以下になる。従って、ループ系列に至るルールの報酬獲得効率 (すなわち価値) を誤って高く評価することを防ぐことができる。

§2 価値を均一に更新できない場合

前項では、状態の混同が起こる観測で、報酬を得るために必要なルールの集合が変化しない場合のみを扱った。しかし、この制限を外すと、提案手法および EPS で報酬獲得に重要なルールの価値を必ずしも均一には更新できない場合が存在する。

まず、決定論的な状態遷移を持つ POMDPs 環境中の複数のルール選択を必要とする観測で、必要とするルールの組み合わせが複数存在し、さらに複数の組み合わせに共通するルールがある場合が考えられる。例えば、ある不完全知覚の影響を受ける観測 o において、行動 a と b の選択を確率的に行うことによって報酬が得られ (片方のみでは得られない)、また行動 a と c の選択を確率的に行うことによって報酬が得られる (片方のみでは得

られない) ような場合である。そのような観測では、複数の組み合わせに共通するルール (a) と共通しないルール (b と c) の間で価値に差が出る可能性がある。特に EPS は累積値ベースの学習法なので、複数の組み合わせに共通するルールと共通しないルールの強化回数の差による影響を受け、両者の価値に大きな差が生じる可能性がある。一方、提案手法は平均値ベースの学習法なので、EPS のように強化回数の差による影響を受けることなく、比較的均一に近い価値が得られることが期待できる。

次に、複数報酬を持つ POMDPs 環境中で、ある報酬を得るために同一観測に必要な複数のルールのうち、一部のルールのみが別の報酬を得るために必要となる場合が考えられる。EPS はそもそも複数報酬には対応していない。提案手法では、学習途中でエピソードごとに異なる報酬への経路を経験した場合、ある報酬を得るために複数のルールを必要とする観測で、別の報酬への経路の影響でその複数のルールの均一性が崩れる可能性がある。しかし、提案手法では学習の進行に伴ってより高い報酬獲得効率を持つ報酬への経路に学習の対象が絞られると考えられるため、学習が進むにつれ、この問題による影響は小さくなると期待できる。ただし、学習途中にこの問題の影響を受けることで、本来報酬獲得効率の高い報酬への経路が過小評価され、本来報酬獲得効率の低い報酬への経路を学習する可能性もあると考えられる。

最後に、確率論的な状態遷移を持つ POMDPs 環境中の学習を考える。このような環境でも、同じ観測 o として知覚される状態 s_1 と s_2 で、有効なルールが異なっていて、 s_1 ではルール (o, b) 、 s_2 ではルール (o, a) を選択すべきであるという図 1 のような状況は、容易に生じ得る。このような場合に、確率論的な状態遷移のおかげで、どちらかのルールのみを選択し続けていても比較的容易にループ (見かけ上のものも含む) から抜けられる場合には、不完全知覚が大きな問題とはならない。問題となるのは、多くのエピソードで両方のルールが選択されているが、まれに片方のルールのみで報酬を得ているような場合で、この場合には、両方のルールを均一に強化することはできない。この場合は決定論的な環境に近いので、均一性は保たれないものの、両方のルールを十分に強化できることが期待できる。ただし、学習途中においては、EPS、提案手法共にルール強化の均一性の崩れから、ループへと陥る可能性がある。EPS は累積値ベースの学習法なので、ルールの価値に対して小さな強化値は学習に影響を与えにくい。そのため、ループに陥ると均一強化時に各ルールが得る強化値が全体的に小さくなるため、政策の改善が難しくなると考えられる。一方、提案手法は平均値ベースの学習法なので、ループに陥った場合には小さな強化値で均一強化が行われ、全体的にルールの価値を下げる形で学習が進む。この際、より高い価値を持つルールほどその価値が下がりやすいため、ルールの価値は均一に近づくと考えられる。従って、EPS に

^{*15} LMS 法 ($f(x) = 1$) などの他の平均値ベースの学習法の強化関数を使うこともできる。ただし、その場合には、複数報酬環境への適応性は保証できない。

図 7 POMDPs 迷路問題

図 8 複数報酬を持つ POMDPs 迷路問題

比べこの問題による影響は小さいと考えられる。

5.2 実験

提案手法の性能を、EPS、RAE-RL との比較実験で確認する。実験環境としては、Sutton の迷路問題に最適ルール分散型の問題が付与された問題環境である図 7 の環境 [植村 05] と、図 7 の環境に複数のゴール報酬を配置した図 8 の環境を用いる。図 7 の環境では、状態 G に到達した際に 10 の報酬が得られる。図 8 の環境では、状態 G_1 に到達した際には 10 の報酬が得られ、状態 G_2 、 G_3 、 G_4 に到達した際には 0.01 の報酬が得られる。これらの環境は、それぞれ真ん中の 2 つの行で横軸のセンサが故障し状態が区別できなくなっている状況を想定している。図中の A 、 B と書かれた状態は不完全知覚の影響を受け、それぞれ同一の観測と見なされる。その他の設定は、4.3 節の実験と同様とする。

以下の全ての実験で、行動選択には正規ボルツマン選択を用い、学習の進行と共に、式(7)に従って温度パラメータ T の値を下げていく。式(7)内のパラメータは、各実験中の手法ごとに試行錯誤により定める。

まず、図 7 に示す環境で実験を行う。この環境では、報酬獲得のために観測 A 、 B でそれぞれ複数の行動が必要となる。観測 A では上と下への移動が必要となり、観測 B では上と下への移動か、或いは上と右への移動が必要となる。観測 B は、前節で述べた価値を均一に更新できない場合の「必要とするルールの組み合わせが複数存在」する場

表 3 POMDPs 迷路問題の観測 A における行動選択確率

行動 \ 手法	提案手法	RAE-RL	EPS
上	0.499954	0.706	0.456
下	0.499954	0.263	0.456
左	0.000002	0.011	0.036
右	0.000091	0.020	0.052

合に相当する。実験環境において、観測 A 、 B で上下への移動を半分の確率で選択し、その他の観測では最適行動を取る政策を実行したところ、平均性能として 0.357 が得られた。また、観測 A で上下への移動を、観測 B では上右への移動をそれぞれ半分の確率で選択し、その他の観測においては最適行動を取る政策を実行したところ、平均性能として 0.337 が得られた。最適政策は簡単にはわからないので、これらの性能値を目安とする。式(7)内のパラメータは、提案手法では $T_s = 0.5$ 、 $\Gamma = 0.9999$ 、 $T_m = 0.01$ 、EPS では $T_s = 1.0$ 、 $\Gamma = 0.9999$ 、 $T_m = 0.01$ 、RAE-RL では $T_s = 0.5$ 、 $\Gamma = 0.9999$ 、 $T_m = 0.05$ とした。

図 7 の環境における実験結果を図 9 に示す。この実験結果から、提案手法の POMDPs 環境の問題に対する学習性能の優越性が確認できる^{*16}。実験終了時の観測 A 、 B での各学習法の行動選択確率をそれぞれ表 3、表 4 に示す。観測 A において、RAE-RL では上行動の選択確率が下行動の選択確率より高くなっているのに対し、提案手法では EPS と同様に上下行動を均一強化している。ここから、提案手法では、最適ルール分散型の問題への頑健性が RAE-RL よりも向上していることが分かる。なお、表 3 から EPS は最適ルール分散型の問題に対して RAE-RL 以上の頑健性があると考えられるが、収束解は RAE-RL の方が高い結果となった。EPS ではループ系列上のルールは全て均一に強化されるため、最適ルール分散型の問題への頑健性とは関係ないルールも均一に強化される。この問題環境では、エピソードのほとんどのルールが見かけ上のループ系列内に含まれるため、最適ルール分散型の問題への頑健性とは関係ないルールの報酬に対する距離情報をうまく扱うことができず、局所解に陥り学習性能の低下を招いたと考えられる。提案手法では、最適ルール分散型の問題への頑健性を得るための必要最小限の均一強化を行うことで、この点を改善している。また表 4 では、観測 B でルール強化の均一性が崩れていることが示されている。ただし、上と右のルールはほぼ同等な選択確率となっており、前節で予想した通り、「必要とするルールの組み合わせが複数存在」する場合にも、十分対応しているといえる。

次に、図 8 の環境で実験を行う。この環境では、ゴール状態 G_1 へ到達するためには図 7 の環境同様観測 A 、 B で複数の行動を選択する必要がある。図 7 の環境において、

*16 この問題と同一の問題において、EPS が Sarsa(λ) より有効であることが示されている [植村 05]。従ってこの問題において、提案手法も Sarsa(λ) より有効であるといえる。

図 9 POMDPs 迷路問題における実験結果

表 4 POMDPs 迷路問題の観測 B における行動選択確率

行動 \ 手法	提案手法	RAE-RL	EPS
上	0.5001846	0.4982	0.419
下	0.0303072	0.0036	0.295
左	0.0000002	0.0009	0.042
右	0.4695080	0.4973	0.244

観測 A, B で上下行動を等確率で選択し他の観測で最適行動を選択した場合, 平均性能として 0.357 が得られた. 図 8 の環境においてもゴール状態 G_1 への最適経路は同じであるため, 同一の性能値となる. これを目安とする. 状態 G_2, G_3, G_4 への最短経路では, 観測 A, B で複数行動を選択する必要が無く, 経路長はそれぞれ 2, 3, 11 となる. すなわち, G_2, G_3, G_4 への最短経路を選択するとそれぞれ $0.01/2 = 0.005$, $0.01/3 \simeq 0.003$, $0.01/11 \simeq 0.001$ の性能値が得られると考えられる. 従って, G_1 への経路を獲得することが, 圧倒的に有用である. 式(7)内のパラメータは, 提案手法では $T_s = 3.0$, $\Gamma = 0.9998$, $T_m = 0.05$, EPS では $T_s = 1.0$, $\Gamma = 0.9999$, $T_m = 0.01$, RAE-RL では $T_s = 1.0$, $\Gamma = 0.9999$, $T_m = 0.05$ とした.

図 8 の環境における実験結果を図 10 に示す. 提案手法が, EPS, RAE-RL と比べて高い性能を得ていることが分かる. 実験終了時の各学習法の観測 A, B での行動選択確率をそれぞれ表 5, 表 6 に示す. EPS ではランダム選択に劣る性能になっている. また, 状態 G_2 や状態 G_3 への経路が学習された結果, 観測 A では上行動を, 観測 B では下行動をほぼ 1 の確率で選択している. そのため, 複数報酬環境にうまく対応できていない. RAE-RL は, 報酬獲得効率に基づいているので, 複数報酬環境に対応している. しかし, 最適ルール分散型の問題に対しては, 表 5 に示すように, 観測 A で上下行動の選択に大きく偏りが生じていて, これが性能の低下につながっていると思われる. 一方, 提案手法では観測 A において上下行動はほぼ同等な選択確率となっている. ただし, 表 3 とは違い完全に均一には強化されていない. これは, 学

図 10 複数報酬を持つ POMDPs 迷路問題における実験結果

表 5 複数報酬を持つ POMDPs 迷路問題の観測 A における行動選択確率

行動 \ 手法	提案手法	RAE-RL	EPS
上	0.317	0.789	0.99986
下	0.363	0.201	0.00005
左	0.133	0.005	0.00005
右	0.188	0.006	0.00005

習途中でたまたま状態 G_2, G_3, G_4 へ到達した (捕まった) 際に均一性が保たれなくなったと考えられる. とはいえ, 選択確率の差は小さく, 前節で予想した通り, 複数報酬を持つ POMDPs 環境にも, 十分対応しているといえる. また表 6 では, 報酬獲得に重要なルールのうち, 一部が突出して強化されるという事態には陥っていないため, 前節で予想した通り, 「必要とするルールの組み合わせが複数存在」する場合にも, 十分対応していることを示している.

最後に, さらに現実的な問題として, 図 8 の環境で確率論的な状態遷移を行う問題を扱った. これは, 5.1.2 節で述べた均一更新を妨げる 3 つの条件全てに当てはまる困難な問題である. 状態遷移は, 行動によって向かった先の正面に確率 0.7, その他の 3 方向に 0.1 ずつで遷移する (ただし, 壁や環境の外への遷移となる場合は, その場に留まる) とした. 実験環境において, 観測 A, B で上下への移動を半分の確率で選択し, その他の観測では最適行動を取る政策を実行したところ, 平均性能として 0.198 が得られた. この性能値は状態 G_2, G_3, G_4 へ到達した際の報酬 (0.01) よりも大きいので, 明らかに G_1 への経路を獲得す

表 6 複数報酬を持つ POMDPs 迷路問題の観測 B における行動選択確率

行動 \ 手法	提案手法	RAE-RL	EPS
上	0.379	0.461	0.00007
下	0.281	0.095	0.99975
左	0.059	0.002	0.00010
右	0.281	0.442	0.00007

図 11 確率論的な複数報酬の POMDPs 迷路問題における実験結果

ることが有用であるといえる．最適政策は簡単にはわからないので，この性能値を目安とする．式(7)内のパラメータは，提案手法では $T_s = 3.0$, $\Gamma = 0.9998$, $T_m = 0.05$, EPS では $T_s = 3.0$, $\Gamma = 0.9995$, $T_m = 0.05$, RAE-RL では $T_s = 0.5$, $\Gamma = 0.9999$, $T_m = 0.05$ とした．

実験結果を図 11 に示す．提案手法が，EPS , RAE-RL と比べて高い性能を得ていることが分かる．ただし，問題の難易度が増したため，目安となる性能と比べて，前の 2 つの実験ほどうまく対応できていない．しかし従来手法と比べると，「必要とするルールの組み合わせが複数存在」する観測を持ち，複数報酬と確率的な状態遷移を持つ POMDPs 環境において，より適切な政策を獲得しているといえる．

なお，詳細は割愛するが，図 7 の環境に同様の確率論的な状態遷移を加えた問題でも性能比較を行った．この確率論的な単独報酬の POMDPs 環境においても，提案手法は EPS , RAE-RL を上回る性能に収束した（総ステップ数 50000 の段階で，提案手法：0.103 , EPS : 0.0305 , RAE-RL : 0.0869）．

以上に示した通り，提案手法は，単独報酬および複数報酬の両方の POMDPs 環境において，EPS , RAE-RL よりも有効であるといえる．

6. ま と め

本論文では，複数報酬環境に適応性のある学習法である RAE-RL に着目し，強化学習法を報酬の分配法と価値の計算法の観点から分類した．そして，強化値の平均を価値として用いる強化学習法の持つ性質を分析し，その学習法が，POMDPs 環境中での学習に有効であることを示した．

また，POMDPs 環境中で適切に学習するためには，均一更新という条件が重要であることに着目し，POMDPs 環境に適した報酬獲得効率に基づく強化学習法を提案した．そして，提案手法が，不完全知覚と複数報酬という

2 つの問題のそれぞれにうまく対応していることを，実験により示した．

平均値ベースの学習法が持つループ系列に対する抑制効果は，実験によって確認はされたものの，ルールの期待価値の算出が複雑なため，定量的な証明は今後の課題である．また，マルチエージェント環境などのさらに現実的な POMDPs 環境中での学習への適用性も検証する必要がある．

◇ 参 考 文 献 ◇

- [Arai 01] Arai, S. and Sycara, K.: Credit Assignment Method for Learning Effective Stochastic Policies in Uncertain Domains, in *Proc. the Genetic and Evolutionary Computation Conference (GECCO-2001)*, pp. 815–822 (2001)
- [Grefenstette 88] Grefenstette, J. J.: Credit Assignment in Rule Discovery Systems Based on Genetic Algorithms, *Machine Learning*, Vol. 3, pp. 225–245 (1988)
- [石村 01] 石村 健二, 天正 新二郎, 山口 智浩: RAE-PIA: 複数報酬環境下における最適政策の効率の強化学習, 第 15 回人工知能学会全国大会論文集 (2001)
- [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)
- [Loch 98] Loch, J. and Singh, S.: Using Eligibility Traces to Find the Best Memoryless Policy in Partially Observable Markov Decision Processes, in *Proc. 15th International Conference on Machine Learning (ICML'98)*, pp. 323–331 (1998)
- [Mahadevan 96] Mahadevan, S.: Average Reward Reinforcement Learning: Foundations, Algorithms, and Empirical Results, *Machine Learning*, Vol. 22, No. 1–3, pp. 159–195 (1996)
- [松井 03] 松井 藤五郎, 犬塚 信博, 世木 博久: POMDPs への行動優先度学習型強化学習アルゴリズムの適用, 第 17 回人工知能学会全国大会論文集 (2003)
- [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ての理論的考察, 人工知能学会誌, Vol. 9, No. 4, pp. 580–587 (1994)
- [宮崎 99] 宮崎 和光, 荒井 幸代, 小林 重信: POMDPs 環境下での決定的政策の学習, 人工知能学会誌, Vol. 14, No. 1, pp. 148–156 (1999)
- [宮崎 03] 宮崎 和光, 小林 重信: Profit Sharing の不完全知覚環境下への拡張: PS-r* の提案と評価, 人工知能学会論文集, Vol. 18, No. 5, pp. 286–296 (2003)
- [Russell 95] Russell, S. J. and Norvig, P.: *Artificial Intelligence: A Modern Approach*, chapter 20, Prentice Hall (1995)
- [斎藤 04] 斎藤 健, 増田 士朗: 不完全知覚判定法を導入した Profit Sharing, 人工知能学会論文集, Vol. 19, No. 5, pp. 379–388 (2004)
- [Singh 94] Singh, S. P., Jaakkola, T., and Jordan, M. I.: Learning Without State-Estimation in Partially Observable Markovian Decision Processes, in *Proc. 11th International Conference on Machine Learning (ICML'94)*, pp. 284–292 (1994)
- [Singh 96] Singh, S. P. and Sutton, R. S.: Reinforcement Learning with Replacing Eligibility Traces, *Machine Learning*, Vol. 22, No. 1–3, pp. 123–158 (1996)
- [Sutton 90] Sutton, R. S.: Integrated Architectures for Learning, Planning, and Reacting Based on Approximating Dynamic Programming, in *Proc. 7th International Conference on Machine Learning (ICML'90)*, pp. 216–224 (1990)
- [高玉 03] 高玉 圭樹: マルチエージェント学習 相互作用の謎に迫る, 第 3 章, コロナ社 (2003)
- [内田 99] 内田 圭, 渡邊 亮介, 山口 智浩: Self-Rewarded RAE:

学習の合理性と高速性とを両立させた経験強化型学習法, 第 13 回人工知能学会全国大会論文集, pp. 281–283 (1999)

[植村 05] 植村 渉, 上野 敦志, 辰巳 昭治: POMDPs 環境のためのエピソード強化型強化学習法, 電子情報通信学会論文誌, Vol. J88-A, No. 6, pp. 761–774 (2005)

[Watkins 92] Watkins, C. J. C. H. and Dayan, P.: Technical Note: Q-learning, *Machine Learning*, Vol. 8, No. 3–4, pp. 279–292 (1992)

[山口 00] 山口 智浩, 石村 健二: RAE-PIA: 報酬獲得効率を最大化する政策の強化学習, 第 14 回人工知能学会全国大会論文集, pp. 329–332 (2000)

〔担当委員: 山口 智浩〕

2007 年 6 月 20 日 受理

著 者 紹 介

河合 宏和(正会員)

2005 年 3 月大阪市立大学工学部情報工学科卒業。2007 年 3 月同大学大学院工学研究科電子情報系専攻修士課程修了。2007 年 4 月株式会社ダイヘン入社, 現在に至る。修士(工学)。

上野 敦志(正会員)

1991 年 4 月東京大学工学部航空学科卒業。1996 年 3 月同大学大学院航空宇宙工学専攻博士課程単位取得退学。1996 年 4 月奈良先端大学情報科学研究科助手。2004 年 4 月大阪市立大学工学部情報工学科講師, 現在に至る。博士(工学)。

辰巳 昭治(正会員)

1970 年 3 月大阪大学工学部通信工学科卒業。1972 年 3 月同大学院工学研究科通信工学専攻修士課程修了。1972 年 4 月川崎重工業(株)入社。1978 年 3 月大阪大学大学院工学研究科通信工学専攻博士課程修了。工学博士。豊橋技術科学大学を経て, 現在, 大阪市立大学工学部情報工学科教授。