論 文 Technical Papers

Profit Sharingの不完全知覚環境下への拡張：PS-r*の提案と評価

An Extention of Profit Sharing to Partially Observable Markov Decision Processes : Proposition of PS-r* and its Evaluation

宮崎 和光 Kazuteru Miyazaki
大学評価・学位授与機構 学位審査研究部
Faculty of Assessment and Reserach of Degrees, National Institution for Academic Degrees
teru@niad.ac.jp, http://svrrd2.niad.ac.jp/faculty/teru/indexj.html

小林 重信 Shigenobu Kobayashi
東京工業大学 大学院総合理工学研究科
Graduate School of Interdisciplinary Science and Engineering, Tokyo Institute of Technology
kobayasi@dis.titech.ac.jp

keywords: reinforcement learning, profit sharing, rational policy making algorithm, POMDPs, theorem

Summary

We know *the rationality theorem of Profit Sharing*(PS) [Miyazaki 94, Miyazaki 99b] and *the Rational Policy Making algorithm*(RPM) [Miyazaki 99a] to guarantee the rationality in a typical class of Partially Observable Markov Decision Processes (POMDPs). In this paper, we focus on the whole class of POMDPs and propose *PS-r* that is an algorithm connected PS and RPM with random selection. In the first, we have analyzed the behavior of PS-r. We have derived that the maximum value of the step to get a reward by PS-r divided by that of random selection is $(r\frac{(1+\frac{M-1}{r})^n}{M^n})$ where n is the maximum number of state that senses same state due to the agent's sensory limitation and M is the number of actions. Furthermore, we propose *PS-r** that can improve the behavior of PS-r. Through numerical examples, we conform the effectiveness of PS-r*.

1. はじめに

強化学習とは, 報酬という特別の入力を手がかりに環境に適応する機械学習システムである. 明示的に正解を与える教師なしに学習できるので魅力的な枠組みと言える. 著者らは, 強化学習研究を**環境同定型**と**経験強化型**のふたつに類別している [宮崎 95]. 環境同定型とは, **離散マルコフ決定過程** (MDPs) で記述される環境を同定し, そこでの最適性を保証する接近である. 現在, 環境の状態遷移確率を明に同定するものとして **k-確実探査法** [宮崎 95], 暗に同定するものとして **Q-learning**(QL) [Watkins 92] が知られている.

環境同定型は MDPs 環境下で最適性が保証される半面, 膨大な試行錯誤を要するという欠点をもつ. そこで, 最適性を重視せず, 経験した範囲内での合理性を保証する接近が経験強化型である. 現在, 経験を系列すなわちエピソードの形で記憶し学習するものとして **Profit Sharing**(PS) [Grefenstette 88, 宮崎 94, 宮崎 99b] や **モンテカルロ法** [Sutton 98], 状態遷移の有無のみを記憶し学習するものとして**合理的政策形成アルゴリズム** (RPM) [宮崎 99a] が知られている.

一般に, 経験強化型は, 環境同定型に比べ学習が早いという特長を持つ. また, 非マルコフ的環境に対する頑健性も重要なポイントとされる. 現在, 非マルコフ的環境の代表に**部分観測マルコフ決定過程** (POMDPs) が知られている. POMDPs とは, 実際には異なる環境の状態が学習器にとっては同一の感覚入力として知覚される, すなわち**不完全知覚状態** [Chrisman 92] を有する問題クラスのことをいう.

POMDPs に対する伝統的な接近法は過去の履歴を用いて不完全知覚状態を分離する**メモリーベース法** [Chrisman 92, McCallum 93, McCallum 95, 末松 98] である. メモリーベース法は, 不完全知覚状態の分離という直接的な方法で POMDPs に対峙する接近である. しかしそのためには, 通常, 状態数の指数オーダーにものぼる膨大な量のメモリーが必要とされる.

そのようなメモリーベース法がもつ欠点を克服するために, **確率的政策** [Singh 94] が提案されている. そこでは, 各感覚入力に対し, 複数の行動の中からひとつの行動を確率的に選択することで, 不完全知覚状態からの脱出

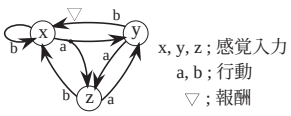

図 1 用語の定義で用いた環境.

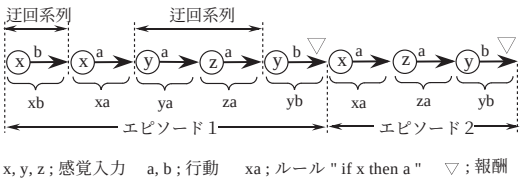

図 2 エピソードと迂回系列の例.

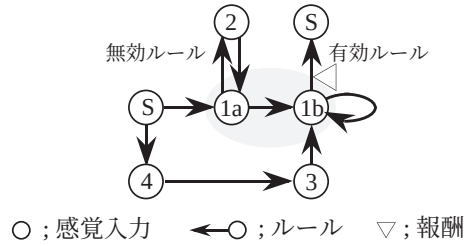

図 3 タイプ 2 の混同の例.

を試みる. 最も簡単で, かつ, 確実に報酬を得ることができる確率的政策は, つねにすべての行動を等確率で選択する**ランダム選択**である. それに対し, 確率的政策を学習する従来手法 [Williams 92, Jaakkola 94, 木村 96, 林 99, Sutton 00, Konda 00, Aberdeen 01] は, 一種の山登り法である. 山登り法は, 局所的な最適性を保証することはできるが, 2.3 節で述べるように, ランダム選択を改善できる可能性がある場合に, 必ずしもそれを改善できるとは限らず, また, 改悪することも起こり得る.

著者らは, 先に, ある特定の POMDPs 環境における, PS および RPM の合理性を保証する定理を導出した [宮崎 94, 宮崎 99a, 宮崎 99b]. しかしそこでは, 確率的政策が要求される一般の POMDPs 環境については議論されていない. そこで本論文では, まず初めに, PS および RPM に確率的政策の考えを導入した手法を提案し, 一般の POMDPs 環境における挙動を解析する. さらに従来のメモリーベース法よりも少ないメモリで不完全知覚状態を分離する手法を提案し, 上記の手法と組み合わせることで, ランダム選択を積極的に改善する手法を提案する.

以下, 第 2 章では, 本論文で使用する用語の定義, 対象問題クラスならびに POMDPs に対する従来手法について述べる. 第 3 章では, PS および RPM に確率的政策の考えを導入した手法である **PS-r** を提案し, 一般の POMDPs 環境下での挙動を解析するとともに **PS-r*** と呼ばれる PS-r の改善手法を提案する. 第 4 章は数値例であり, 提案手法の有効性を確認する. 第 5 章は結論であり, 本研究の成果を総括し, 今後の課題をとりまとめる.

2. 問 題 設 定

2·1 準 備

本論文では, POMDPs で記述される環境を対象とする. 学習器は環境からの**感覚入力**に対し, **行動**を選択し, 実行に移す. 一連の行動に対して, 環境から**報酬**が与えられる. 以下では, 正の報酬のみを扱い, 負の報酬である**罰**の存在は考えない. 時間は認識-行動サイクルを 1 単位として離散化される. 感覚入力は離散的な属性-値ベクトルとして与えられ, 行動は M 個の離散的なバリエーションの中から選ばれる. 以下では, 感覚入力の状態を単に**状態**と呼ぶ. ある状態において実行可能な行動は**ルール**として記述される. 状態 x で行動 a を選択する"if x then a" というルールを xa と書く. 各状態に対し, 選択すべき行動を与える関数を**政策**と呼び, 単位行動当たりの期待獲得報酬量が正である政策を**合理的政策**, それを最大化する政策を**最適政策**と呼ぶ.

初期状態あるいは報酬を得た直後から次の報酬までのルール系列を**エピソード**という. 例えば図 1 の環境で学習器が [xb, xa, ya, za, yb, (報酬), xa, za, yb, (報酬)] と行動したとすると, このなかには (xb, xa, ya, za, yb), (xa, za, yb) のふたつのエピソードが含まれている (図 2 参照). あるエピソードで, 同一の状態に対して異なるルールが選択されているとき, その間のルール系列を**迂回系列**という. 例えば図 2 のエピソード 1 には (xb), (ya, za) のふたつの迂回系列が含まれている. 現在までのすべてのエピソードで, つねに迂回系列上にあるルールを**無効ルール**と呼び. それ以外を**有効ルール**と呼ぶ.

不完全知覚により, 有効ルールと無効ルールが同一視されることを**タイプ 2 の混同**という [宮崎 99a][*1]. 例えば, 図 3 の環境を考える. この環境では, 状態 1a で上という行動は無効ルールであるが, 同じ行動は状態 1b では有効ルールとなる. 学習器は状態 1a と 1b をともに同じ状態 (状態 1) と認識するため, 状態 1 で上という行動は, 学習器にとっては有効ルールとされる. しかし, たとえそのルールを選んだとしても, 状態 S で右へ向かう行動を学習した場合には, 状態 1a と 2 の間を往復する合理的でない政策が学習される.

2·2 PS に基づく強化学習

著者らが行ってきた Profit Sharing (PS) に基づく強化学習研究は, 図 4 のようにまとめられる.

---

*1 なお, 状態の価値が混同されることを**タイプ 1 の混同**という. タイプ 1 の混同は,QL では対処できないが PS では問題とならない [宮崎 99a].

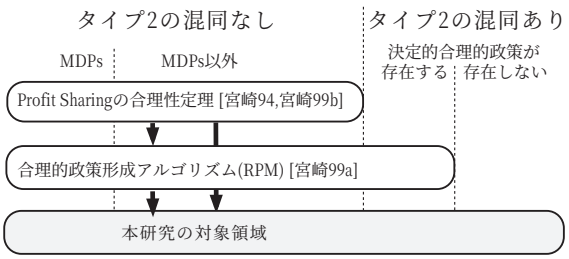

図 4 本研究の位置付け.

§1 PS の合理性定理

PS とは, 報酬を得たときにそのとき経験したエピソードを一括的に強化する手法である. PS ではエピソード単位でルールに付加された評価値を強化する. 報酬からどれだけ過去かを引き数とし, 強化値を返す関数を**強化関数**と呼ぶ.

タイプ 2 の混同が存在しない環境下で, 無効ルールと有効ルールとが競合するならば, 明らかに無効ルールを強化すべきではない. そのような環境下で任意の無効ルールを抑制し報酬を得つづけるための強化関数の必要十分条件が [宮崎 94] により示されている. 以下では, この定理を **PS の合理性定理**と呼ぶ.

§2 合理的政策形成アルゴリズム

有効ルールの定義より, あるエピソードにおいて同一の感覚入力に対する行動選択の中で報酬に最も近い位置で選択された行動を含むルールは有効ルールである. この性質を利用し, 合理的政策のより効率的な獲得を目指した手法が図 5 に示す**合理的政策形成アルゴリズム** (RPM) [宮崎 99a] である.

```
procedure RPM
 begin
  do
   1 次および 2 次記憶領域の内容を初期化する.
   do
     感覚入力xを知覚する.
     if xに対応する 2 次記憶上に行動が記憶されている then
       その行動を選択する.
     else 環境探査戦略により行動aを選択し,
       その行動を 1 次記憶上に上書きする.
     if 報酬を得た then 1 次記憶領域の
       内容を 2 次記憶領域に複写する.
   while ( 2 次記憶領域が未収束)
   if 合理的政策が得られている then 2 次記憶領域の内容を
     保存する.
  while /* マルチスタート法 */
 end.
```


図 5 RPM のアルゴリズム.

RPM では, まず, **1 次**および **2 次記憶領域**と呼ばれる状態数長の 1 次元配列を用意する. 行動を選択するごとに, 1 次記憶上に, そのとき選択した行動を上書きする. 例えば, 状態 x で行動 a を選択した際は, 1 次記憶上の状態 x に行動 a を上書きする. 以上を報酬が得られるまで繰り返し, 報酬が得られた時点で,1 次記憶領域の内容を 2 次記憶領域に複写する. この結果,2 次記憶領域には, 有効ルールのみが記録されていく. 学習器は, 有効ルールが判明している感覚入力を知覚した場合には, そのルールに従って行動を出力し, そうでない場合には, 環境を探査するための行動を出力する.

2 次記憶には政策が保存されるので, その収束を判定すれば, 学習の打ち切りが可能となる. 収束の判定としては, 2 次記憶が最後に更新されたステップ数を n とし, その 2 倍の $2n$ ステップ行動を出力しても 2 次記憶が更新されなければ, 収束したとみなす方法が例えば考えられる. タイプ 2 の混同が存在しないクラスでは, 上記の収束判定後, 必ず合理的政策が得られる. しかし, タイプ 2 の混同が存在する場合には, 得られた政策が合理的政策になっているとは限らない. 収束判定期間内に報酬が全く得られなければ, 合理的政策の学習に失敗したと判断される. この場合,1 次および 2 次記憶の内容を初期化して, 新しい政策の学習を行うものとする. このような新しいエピソードを任意に生成し, 政策の形成をやり直す手法を**マルチスタート法**と呼んでいる.

PS の合理性定理は, タイプ 2 の混同が存在しないことを要請するが, RPM は決定的な合理的政策が存在すればそれを学習することができる. そのため RPM は, 図 4 に示すように PS の合理性定理よりも, より広いクラスを対象としていると言える.

2·3 従来の POMDPs に対する接近

本節では, 前節で述べた **PS に基づく強化学習**以外の POMDPs に対する接近法について概説する.

§1 メモリーベース法による接近

メモリーベース法は, 過去の履歴を用いて, 不完全知覚状態を分離する接近である [Chrisman 92, McCallum 93, McCallum 95, 末松 98]. 不完全知覚状態の分離には, 通常, 統計的検定が利用されるが, 従来手法では, 有意な結果を得るためには過去の膨大な量の履歴をメモリーに保存しておく必要がある. 代表的なメモリーベース法として **Utile Suffix Memory** (USM) [McCallum 95] が知られている.

USM では, 過去の履歴を木構造で表現し, 各葉ノードを内部状態とすることにより, 可変長の履歴を扱う. ここで参照すべき過去の履歴の長さは, **fringe** というパラメータで制御される. fringe が短すぎると, 不完全知覚状態を正しく分離できない場合がある. そのため fringe は十分長くとる必要があるが, USM が要する記憶容量は, 学習器が観測する状態の種類を N とすれば, 行動の種類を無視したとしても最悪の場合, $O(N^{fringe})$ となり, 状態数の指数オーダーにものぼる.

USM と同様に, 木構造の可変長の履歴を用いる手法に **BLHT** [末松 98] がある. BLHT は, ベイズ統計を利用し, USM よりも効率よく POMDPs モデルの学習を行う手法である. しかしメモリー量に関する問題が解決されているとは言えない.

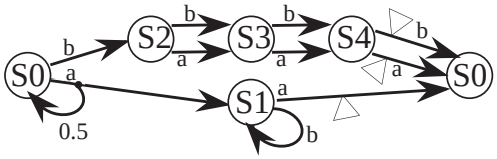

図 6 山登り法の限界を示すために用いた環境.

§2 確率的政策に基づく接近

そのようなメモリーベース法が有する欠点を克服するために, **確率的政策** [Singh 94] が提案されている. そこでは, 各感覚入力に対し, 複数の行動の中からひとつの行動を確率的に選択することで, 不完全知覚状態からの脱出を試みる.

最も簡単で, かつ, 確実に報酬を得ることができる確率的政策はランダム選択である. それに対し, 確率的政策を学習する従来手法 [Williams 92, Jaakkola 94, 木村 96, 林 99, Sutton 00, Konda 00, Aberdeen 01] では, 一種の山登り法により政策を改善していく. 特に [木村 96] の**確率的傾斜法**をはじめ [Sutton 00, Konda 00, Aberdeen 01] らの手法では, 今までとった行動の情報論的な意味での価値を, **適正度の履歴** (eligibility trace)[Sutton 98][*2]という形で過去の行動ほど割り引いて記憶していくことで, 不完全知覚状態に対応している. 山登り法は, 局所的な最適性を保証することはできるが, ランダム選択を改善できる可能性がある場合に, 必ずしもそれを改善できるとは限らず, また, 改悪することも起こり得る.

例えば, 図 6 の環境を考える. この環境で, ランダム選択が報酬を得るために要するステップ数の期待値は 4 である. 最適政策は状態 $S0$ および状態 $S1$ で行動 a を選択した場合であり, その場合の期待ステップ数は 3 である. 一方, 状態 $S0$ で行動 b を選択した場合には, ランダム選択同様, 報酬を得るためには 4 ステップ必要となる. 実際に, この環境に対し, ランダム選択を [木村 96] の確率的傾斜法を用い改善させたところ, 乱数の種を変えて行った 100 回の実験にのおける報酬獲得に要した平均ステップ数は 3.78 であったが, ランダム選択よりも改善されていたのは, 100 回の実験中 25 回のみであった. それ以外の 73 回は, ランダム選択と同等の期待値 4 の政策が学習され, 残る 2 回はランダム選択よりも改悪される結果となった. [*3]

§3 その他の接近

上記以外の接近法として, [Wiering 96] は **Levin's Serach** と呼ばれる全探索法の一種を用いて POMDPs 問題に対処する手法を提案している. しかし, そこでは, 状態遷移が決定的でなければならないという重大な制約がある.

また,[Lin 92, Whitehead 95, Glickman 01] らはリカレントニューラルネットワークを利用することで POMDPs 問題に対処する手法を提案している. しかし, ニューラルネットワークを用いるため学習結果はブラックボックス的である.

2·4 本論文の接近法

POMDPs 環境下では, ランダム選択を用いれば必ず報酬を得ることができる. しかし, 従来の確率的政策を学習する手法は, ランダム選択を改善できる可能性がある場合に, 必ずしもそれを改善できるとは限らない. 一方, 従来のメモリーベース法は不完全知覚状態の分離に状態数の指数オーダーという非現実的な量のメモリを要請する. そこで本論文では, 従来のメモリーベース法よりも少ないメモリで不完全知覚状態を分離し, かつ, ランダム選択を積極的に改善することが可能な手法の提案を目指す.

ところで, 合理性定理を満たす PS や RPM は, つねに無効ルールよりも有効ルールを強化する. したがって, 環境中にタイプ 2 の混同が存在しなければ, 報酬獲得までの行動数がランダム選択より多くなることはない. 一方, タイプ 2 の混同が存在する場合には, 無効ルールを選択しなければ報酬が得られない場合があるので, これらの手法では, 報酬獲得までの行動数がランダム選択より多くなる可能性がある.

そこで本論文では, まず初めに, PS および RPM に確率的政策の考えを導入した手法を提案し, 一般の POMDPs 環境における挙動をランダム選択との比較の下で解析する. そして, その結果を踏まえ, 従来のメモリーベース法よりも少ないメモリでランダム選択を積極的に改善する手法の提案を行う.

3. PS の不完全知覚環境下への拡張

3·1 PS-r の 提 案

PS や RPM をタイプ 2 の混同が存在する環境下に拡張することを考える. そのために, まず, PS や RPM に確率的政策の考えを導入した手法である **PS-r** を提案する.

図 7 に PS-r のアルゴリズムを示す. まず, 全ルールは無効ルールに初期化され, r $(0 < r < 1)$ 点が付与される. r の初期値は任意である. PS-r の行動選択は, 学習時はランダム選択, 学習結果を評価する際には, r の値に基づくルーレット選択に従うものとする.

エピソードごとに,RPM と同様な手法で, 有効ルールが判定される. まず **1 次記憶領域**と呼ばれる状態数長の 1 次元配列を用意する. 学習器は行動を出力するごとに, 1 次記憶上の対応する部分に, そのとき出力した行動を上書きする. 以上を報酬が得られるまで繰り返す. 報酬が得られた時点で, 1 次記憶上に存在しているルールは有

*2 他にも適正度の履歴を用いた手法として, TD(λ), Sarsa(λ), Q(λ) などが知られている [Sutton 98]. 但し, POMDPs 環境に対する有効性は経験的, あるいは限定された条件下 [Loch 98] で示されているのみである.

*3 本来, 理論上は, 改悪は生じないはずである. しかし実際には, 理論が要請している仮定が満たされず, 改悪される場合がある.

```
procedure PS-r
 begin
  1 次記憶領域の内容を初期化する.
  全ルールを無効ルールに初期化しr点付与する.
  do
   感覚入力xを知覚する.
   if 評価時 then rの値に基づくルーレット選択により
     行動aを選択する.
   else ランダム選択により行動aを選択する.
   1 次記憶上に行動aを上書きする.
   if 報酬を得た then 1 次記憶上に存在しているルール
     の rを1にする.
  while
 end.
```

図 7 PS-r のアルゴリズム.

効ルールである. PS-r は学習時にはランダム選択を行うので, これにより, 原理上, すべての有効ルールの発見が可能となる. 有効ルールであると判定されたルールの r は 1 に変化させる. ひとたび 1 点が付与されたルールの r は, 以後 1 点のままとする. これにより有効ルールと無効ルールを区別することができる.

PS の合理性定理では有効ルールがつねに無効ルールよりも強化される. それに対し PS-r は, 無効ルールに対する有効ルールの強化され具合をつねに $\frac{1}{r}$ に固定した PS であると言える. また, 確率的政策をベースにした RPM と考えることもできる.

3·2 PS-r の 性 質

本節では, POMDPs 環境をタイプ 2 の混同の有無で分類し, 各々のクラスにおける PS-r とランダム選択の関係を明らかにする.

§1 タイプ 2 の混同が存在しない場合の性質

タイプ 2 の混同が存在しない場合,PS-r は以下の性質を持つ.

［定理 3·1］（タイプ 2 の混同が存在しない場合の性質） タイプ 2 の混同が存在しなければ, PS-r が報酬を得るために要する行動数が, ランダム選択の行動数よりも多くなることはない.

証明は付録 A を参照されたい.

ところで, タイプ 2 の混同が存在しない状態に無効ルールが存在すれば, 少なくともその状態に関しては, PS-r はランダム選択よりもよい挙動を示すことが可能となる. したがって, 環境中にそのような状態が多ければ多い程, PS-r の挙動はよりよいものになる.

§2 一般の POMDPs 環境における性質

次に, 一般の POMDPs 環境における PS-r の性質を調べる. そのために, まず, POMDPs 環境の中で, PS-r の報酬獲得までの行動数が, ランダム選択の場合に比べ最も多くなる環境の構造, ならびにその構造の下での PS-r の性質について調べる.

［補題 3·1］（最も困難な環境の構造） PS-r の報酬獲得までの行動数が, ランダム選択の場合に比べ最も多くなる環境の構造は, 図 8 の構造で, 行動 a を出力するルールが 1 個, 行動 b を出力するルールが $M-1$ 個存在する場合である. 以下では, この構造を**構造ω**と呼ぶ.

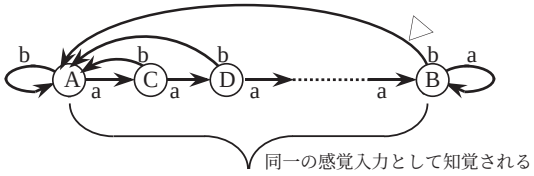

図 8 構造ω.

証明は付録 B を参照されたい.

［補題 3·2］（構造ωにおける比較） 構造ωで, PS-r が報酬を得るまでに要する行動数は, ランダム選択の最大

$$r\frac{(1+\frac{M-1}{r})^n}{M^n}, \tag{1}$$

倍である. ここで n は不完全知覚により同一の状態とみなされている状態の個数の最大値である.

証明は付録 C を参照されたい.

補題 3.1 および 3.2 より, 直ちに, 以下のような一般の POMDPs 環境における定理が導かれる.

［定理 3·2］（一般の POMDPs 環境における性質） PS-r が報酬を得るために要する行動数は, ランダム選択の最大

$$r\frac{(1+\frac{M-1}{r})^n}{M^n}, \tag{2}$$

倍である. ここで n は, 不完全知覚により同一の状態とみなされている状態の個数の最大値である.

$M=2$ のときの (2) 式の値は, 表 1 のようになる. r を 1 に近づければ近づけるほどランダム選択に近づき, $r=1$ の場合はランダム選択に完全に一致する.

表 1 定理 3.2 の計算例 ($M=2$).

n	2	3	4	5
$0.5*1.5^n$ if r=0.5	1.125	1.687	2.531	3.796
$0.9*(19/18)^n$ if r=0.9	1.002	1.058	1.117	1.179

定理 3.2 は, 最悪ケース, すなわち環境が構造ωのみで構成されている場合を想定し導出された定理である. したがって, タイプ 2 の混同が存在していたとしても, それが構造ωでなければ, PS-r の挙動はよりよいものになる. また, 既に述べたように, 環境の一部にタイプ 2 の混同が存在せず, かつ, そのような状態に無効ルールが存在すればする程, PS-r の挙動はよりよいものになる.

3·3 PS-r の 改 良

PS-r は一般の POMDPs 環境下では, ランダム選択に劣る可能性がある. そこで, 次に PS-r を改良し, ランダム選択に劣ることがなく, かつ, 改善できるときには, 積極的に改善を図る手法を提案する.

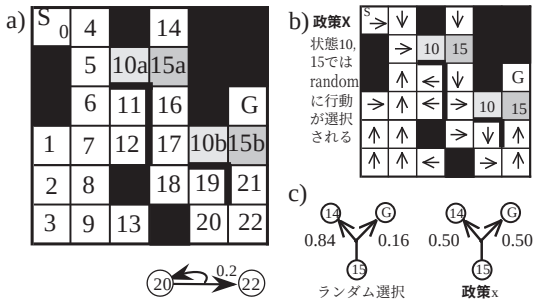

図 9 ある特定の政策 (政策 x) とランダム選択との比較.

POMDPs 環境下, 特に不完全知覚状態下では確率的政策が重要となる. 最も簡単で, かつ, 確実な確率的政策はランダム選択であるが, 既に繰り返し述べているように, 従来の確率的政策を学習する手法は, 必ずしもランダム選択に対する優位性が保証されている訳ではない. 一方, ある状態が不完全知覚状態でなければ, その状態では無効ルールを選択する必要はない. そこで本論文では, 状態の不完全知覚性を判定し, 不完全知覚状態ではランダム選択, それ以外の状態では有効ルールを選択することを考える.

上記の実現には, 状態の不完全知覚性の判定が重要となる. ところで環境中に, 不完全知覚が存在しないならばマルコフ性が成立する. マルコフ性が成立するならば, どのような行動系列を採用したとしても, その状態における各ルールの各状態への状態遷移確率は, 各々独立なある一定の値をとる. それに対しマルコフ性が成立しない場合, すなわち不完全知覚が存在する場合には, 各ルールの状態遷移確率は, そのルールを選択するまでの行動系列に依存して変化する可能性がある. 本論文では, この性質を利用し各状態における不完全知覚の有無を判定する.

具体的には, ルールを選択した際の状態遷移先 (状態) の分布をランダム選択に従った場合と**「ある特定の政策」**に従った場合とで比較する. この比較には χ^2 **検定** [楠 98] を利用する. これは, 政策同士を直接比較する検定なので, 木構造などを記憶する従来のメモリーベース法よりも少ないオーダーで検定可能である. 十分なサンプルに基づきあるルールを検定した結果, これらふたつの政策による分布が等しくないと判断されたルールを選択可能とする状態には, 不完全知覚が存在すると考えられる. これは [宮崎 98, Miyazaki 99] において提案された基本アイデアをより洗練化したものである.

例えば図 9 の環境を考える. 学習器は各状態で 4 種類の行動 (上, 下, 左, 右) の中からひとつを選択できるものとする. この環境では, 状態 20 で右という行動のみが非決定的な遷移をする. また, 状態 10a と状態 10b の間, および状態 15a と状態 15b の間には不完全知覚が存在し, それぞれ状態 10 および状態 15 として知覚される. 但し, タイプ 2 の混同が存在するのは状態 15 のみである. この環境で, 例えば, 不完全知覚が存在しない状態である状態

```
procedure PS-r *
 begin
  do
   1次記憶領域, 不完全知覚判定フラグ, 学習時用および評価時用
   のルールの選択回数, 状態遷移回数を初期化する.
   全ルールを無効ルールに初期化しr点付与する.
   感覚入力xを知覚する.
   if 評価時 then 不完全知覚判定フラグを参照し行動を選択する.
   else
     γ =0 (0は学習時, 1は検定時を意味する).
    do
     if γ = 0 then call(学習モード) else call(検定モード ).
    while
  while /* マルチスタート法 */
 end.

procedure 学習モード
 begin
  ランダム選択により行動aを選択する.
  学習時用のルールの選択回数を更新する.
  1次記憶上に行動aを上書きする.
  if 報酬を得た then 1次記憶上に存在しているルールのrを1に
   する.
   if 検定モードへの切替条件を満たす then γ = 1.
   感覚入力xを知覚する.
   学習時用の状態遷移回数を更新する.
 end.

procedure 検定モード
 begin
  if xが未知状態 then call(未知状態に対する例外処理) .
  else r の値に基づくルーレット選択により行動aを選択する.
    検定時用のルールの選択回数を更新する.
    if 報酬を得た then χ2検定を実行する.
     不完全知覚が存在すると判定された状態の不完全知覚判定
     フラグを立て, 存在しないと判定された状態の不完全知覚
     判定フラグを降ろす.
     if 学習モードへの切替条件を満たす then γ = 0.
    感覚入力xを知覚する.
    検定時用の状態遷移回数を更新する.
 end.

procedure 未知状態に対する例外処理
 begin
  ランダム選択により行動aを選択する.
  if 報酬を得た then γ = 0.
    感覚入力xを知覚する.
 end.
```

図 10 PS-r*のアルゴリズム.

20 で右という行動により状態 22 に遷移する確率は, ランダム選択を行った場合, 図 9b に示すある特定の政策 (政策 x) に従った場合ともに 0.2 であることが期待される. 一方, 例えば, 不完全知覚が存在する状態 15 で上という行動による遷移は各々図 9c) のようになり, 一致しない.

次節では, PS-r に χ^2 検定による不完全知覚の判定機能を付加し, ランダム選択を積極的に改善する手法である **PS-r*** を提案する.

3·4 PS-r*の 提 案

図 10 に PS-r*のアルゴリズムを示す. PS-r*は, **強化学習を行うフェーズ**と**学習した結果を評価するフェーズ**とに大きく分けられる. さらに, 強化学習を行うフェーズは, **学習モード**と**検定モード**に二分される.

§1 学習モードと検定モード

アルゴリズムは, 学習モードから開始される. 学習モードでは, 検定モードでランダム選択との比較に使用される「ある特定の政策」の学習が行われる. まず学習開始時には,PS-r 同様, 全ルールは無効ルールに初期化され, 任意の r $(0 < r < 1)$ 点が付与される. そして,PS-r と同様に, ランダム選択により行動が選択され, 報酬を得た後

に有効ルールが判定され, 有効ルールの r が 1 にセットされる.

一方, 検定モードでは, 学習モードで更新された r の値に基づき行動が選択される. そこでは, 未知状態を経験しない限り r の値に基づくルーレット選択により行動が選択され, 未知状態を経験した後には, 次の報酬が得られるまでランダム選択を行う. これが「ある特定の政策」に相当する.

学習モードと検定モードはある一定の条件で切り替えられる. 切替条件としては, 例えば, 報酬獲得回数やルールの選択回数などが用いられる.

§ 2 χ^2 検定による不完全知覚状態の判定

検定モードでは, 報酬を得た後, 各ルール毎に, 学習モード (ランダム選択) での状態遷移と検定モード (ある特定の政策) での状態遷移を χ^2 検定を用いて比較する. 具体的には, 各ルールごとに各ルールの状態遷移先 (状態) の分布が等しいか否かを検定する.

このような検定には, 各ルールの選択回数ならびに各ルールによる状態遷移先状態を記憶するためのメモリが必要となる. したがって, 状態数を N, 行動の種類を M としたとき, $O(MN)$ および $O(MN^2)$ の大きさをもつ記憶領域を学習モードおよび検定モード用に各々独立に用意する. さらに,χ^2 検定の結果である各状態における不完全知覚の有無を記憶するための状態数長の 1 次元配列である**不完全知覚判定フラグ**を用意する. このフラグは, 不完全知覚あり, すなわちフラグを立てた状態に初期化する.

χ^2 検定の結果, ふたつのモードでの状態遷移が一致しないと判断されたルールを選択可能とする状態には不完全知覚が存在する可能性が高い. それに対し, ある状態において選択可能なすべてのルールにおけるふたつのモードでの状態遷移がすべて一致したと判定された状態には不完全知覚が存在しない可能性が高い. そのような χ^2 検定の結果を踏まえ, 不完全知覚判定フラグを更新する.

本来,χ^2 検定を正しく行うためには, 十分なサンプル (行動選択) が必要である. このサンプルの大きさは, 有意水準 α ならびに検出力 $1-\beta$ を設定することで統計理論から導くことができる [楠 98].

§ 3 学習結果の評価とマルチスタート法の活用

最後に, 評価フェーズ, すなわち学習の結果得られた政策を評価する際には, 不完全知覚判定判定フラグを考慮し行動を選択する. 具体的には, 不完全知覚判定フラグが降ろされている, すなわち不完全知覚が存在しないと判定されている状態では $r<1$ であるルールの r を 0 としてルーレット選択[*4]を行い, 不完全知覚が存在すると判断されている状態ではランダム選択を行う.

その結果, すべての状態が不完全知覚状態である場合, PS-r*の挙動は, ランダム選択と完全に一致する. それに対し, 不完全知覚が存在しない状態に無効ルールが存在すればする程, ランダム選択に対する PS-r*の優位性は確実に増大する.

ところで, ランダム選択との比較に用いた「ある特定の政策」によっては, 正しく不完全知覚状態を検出できない可能性がある. そのような場合であっても評価フェーズで報酬が得られるならば問題ないが, そうでない場合, 比較に用いた「ある特定の政策」を変更する必要がある. PS-r*では, この変更には, マルチスタート法を用いる. 具体的には, ランダム選択による報酬獲得行動数の最悪値を記憶しておき, その値の 2 倍の行動を出力したとしても報酬が得られなければ政策を形成し直す手法が, 例えば, 考えられる.

3･5 PS-r*の 性 質

PS-r*は, 不完全知覚が存在すると判断された状態ではランダムに行動を選択する. そのため, すべての状態に不完全知覚が存在する環境が PS-r*にとって最も困難な環境であり, そのような場合,PS-r*の挙動は, ランダム選択と完全に一致する.

一方,PS-r*は, 不完全知覚が存在しないと判断された状態では有効ルールのみを選択する. そのため, 不完全知覚が存在しない状態に無効ルールが多く存在すればする程, ランダム選択に対する優位性が確実に増大する. これは,PS-r や従来の確率的政策を学習する手法にはない PS-r*の重要な特長である.

PS-r では, 学習時には, つねに r の値に基づき行動が選択されるが, PS-r*では, r が各ルールごとに個別に 0 または 1 に更新される. そのため, 学習開始時に設定される r の値は, PS-r では学習の結果得られる政策の質に直接影響を与えるが, PS-r*では検定時に使用される「ある特定の政策」に影響を与えるのみである.

PS-r は $O(MN)$ のメモリで学習可能であるが, PS-r*では $O(MN^2)$ のメモリを必要とする. 検定を行うためメモリが増加しているが, 従来のメモリーベース法よりは少いオーダーである. また, 正しい検定が行われるためには, 統計理論から導かれる必要なサンプル数を満足するような行動選択が, 一般には, 必要である.

4. PS-r と PS-r*の性能評価

4･1 構造ωを含む環境での評価

§ 1 ランダム選択との比較

i. 実験環境の性質

PS-r および PS-r*の性能を図 11 に示す環境を用いて評価する. この環境中の状態 Z には不完全知覚が存在し, 本来異なる状態である Z_a, Z_b, Z_c, Z_d が同一の状態 (状態 Z) として知覚される. また, 状態 Z にはタイプ 2 の混同が存在し, この部分が構造ωに相当する. ルール Xa には非決定性が存在し, 確率 p で状態 S_1, 確率 $1-p$ で状

*4 有効ルールの r はすべて 1 なので, すべての有効ルールを等しい確率で選択するルーレット選択となる.

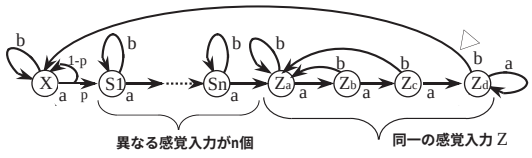

図 11 構造ωを含む環境.

表 2 検定結果に応じた PS-r*の報酬獲得までの行動数の期待値.

		状態 Z	
		未検定	検定成功
状態 X	未検定	-	$\frac{1}{p}+(1+r)n+16$
	検定成功	$\frac{1}{p}+n+\frac{(1+r)^4}{r^3}$	$\frac{1}{p}+n+16$
	検定失敗	$\frac{1}{p}+1+n+\frac{(1+r)^4}{r^3}$	$\frac{1}{p}+1+n+16$

態 X に遷移する. 状態 S_1 から S_n は各々異なる状態として知覚される. n および p の値を調整することで報酬獲得までの行動数が変化する.

この環境で, ランダム選択および PS-r の報酬を得るまでの行動数の期待値は以下のようになる.

- ランダム選択

$$(\frac{1}{p}+1)+(2n)+(16) \qquad (3)$$

- PS-r

$$(\frac{1}{p}+r)+((1+r)*n)+(\frac{(1+r)^4}{r^3}) \qquad (4)$$

ここで第 1 項は状態 X, 第 2 項は状態 $S_i(i=1,2,...n)$, 第 3 項は状態 Z に関する期待値である. これより PS-r は,

$$\frac{1}{1-r}(\frac{(1+r)^4}{r^3}+(r-17)) \qquad (5)$$

より大きい n の範囲ではランダム選択に優るが, (5) 式より小さい n の範囲では, ランダム選択に劣ることがわかる.

PS-r*の政策は, 検定が一度も行われていない間はランダム選択に一致するが, 実際に検定が行われた後は, 状態 X および Z の検定結果に応じ変化する. 状態 Z に関する検定が失敗, すなわち状態 Z に不完全知覚が存在しないと判断した場合には, PS-r*は, ルール Za の選択確率を 0 にしてしまい報酬を得ることができなくなる. それ以外の検定結果は表 2 のようにまとめられる. ここで状態 Z に関する検定が成功とは状態 Z に不完全知覚が存在すると判断したことを意味する. 一方, 状態 X に関しては, 検定成功が状態 X に不完全知覚が存在しないと判断したことを意味し, 検定失敗が状態 X に不完全知覚が存在すると判断したことを意味する. また, 状態 $S_i(i=1,2,...n)$ に関しては, つねに不完全知覚が存在しないと判断できるので表には含めていない.

ところで 3.4 節で述べたように, PS-r*で正しく検定が行われるためには統計理論から導かれる十分な大きさのサンプル (行動選択) が必要である. サンプルの大きさ n

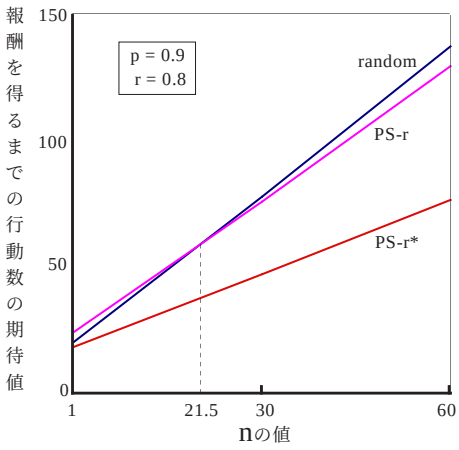

図 12 $p=0.9$, $r=0.8$ とし, n を変えたときの各手法の報酬を得るまでの行動数の期待値.

は, 有意水準を α, 検出力を $1-\beta$ としたとき, 以下の式で与えられる [楠 98].

$$n=\frac{1}{2}(\frac{u(\alpha)+u(2\ \beta)}{sin^{-1}\sqrt{\pi_1}-sin^{-1}\sqrt{\pi_2}})^2 \qquad (6)$$

ここで π_1 および π_2 は, 各々, ランダム選択および「ある特定の政策」に従った場合の検定対象のルールによる, ある注目している状態への状態遷移確率である. また, $u(\cdot)$ は正規分布表から得られる値であり, 例えば $\alpha=0.05$, $\beta=0.10$ の場合, $u(\alpha)=1.960$, $u(2\ \beta)=1.282$ である.

$p=0.9$, $r=0.8$ とし, n を変えたときの, 各手法の報酬を得るまでの行動数の期待値は図 12 のようになる. ここでランダム選択および PS-r に関しては (3) 式および (4) 式の計算結果を示し, PS-r*に関しては状態 X および Z に関する検定が成功した場合の計算結果を示してある. PS-r*は, 検定が正しく行われた後には, 任意の n,p,r に対しランダム選択および PS-r よりも優れていることがわかる. しかし検定成功のためのサンプル数が不足している場合, すなわち検定対象の行動選択回数が (6) 式を満足しない場合には, PS-r*は様々な挙動をとる可能性があるので計算機実験によりその挙動を確認する.

ii. 実験結果および考察

$n=7$, $p=0.9$, $r=0.8$ の場合を対象とし, PS-r*をランダム選択および PS-r と比較する. この環境で, ランダム選択が報酬を得るために要する行動数の期待値は 32.1 である. PS-r は, この環境下では, ひとたび報酬を得れば以後政策は変化しない. それに対し PS-r*は検定結果に応じ学習結果が変化する. また, 最初の報酬を得るまでに要する行動数の期待値は, PS-r, PS-r*ともに 32.1 である. これは, 両手法とも, 最初の報酬を得るまではランダム選択を行うためである.

乱数の種を変えて行った 30 回の実験の結果を図 13 に示す. ここで横軸は行動選択回数, 縦軸はその行動選択回数の時点で得られている政策の質, すなわち, その時点

で得られている政策を用いた場合の報酬を得るまでに要する行動数の平均である．破線がランダム選択の理論値 (32.1) である．また, PS-r*では, 報酬を得るごとに学習モードと検定モードを交互に切り替えた．以下の実験では, すべてこの切り替え条件を用いている．

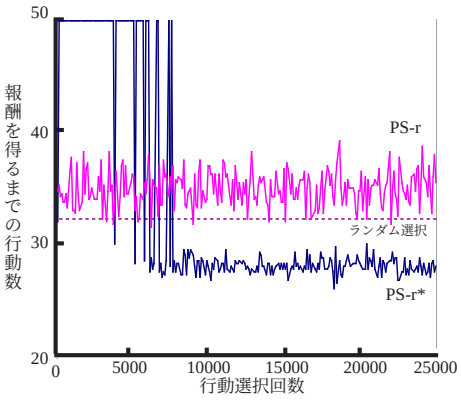

図 13 $n = 7, p = 0.9, r = 0.8$ の場合の結果.

図 13 で PS-r*の結果の一部で縦軸が振り切れているのは, 状態 Z に関する検定が失敗していることを意味する．状態 Z に関する検定が成功した後の微振動は状態 X に関する検定結果の揺れを意味する．一般に, 「不完全知覚が存在しない状態に対する検定誤り」よりも「不完全知覚が存在する状態に対する検定誤り」の方が深刻である．この意味から, 図 11 に示す環境で, 状態 Z に関する検定が先に成功していることは評価に値する．

§2 従来手法との比較

i. メモリーベース法との比較

次に,POMDPs を対象とする代表的なメモリーベース手法である Utile Suffix Memory (USM) [McCallum 95] との比較を行う．

USM は, 図 11 の環境の場合, 状態 Z を区別するために 5 ステップ前に位置していた状態の情報を必要とする．したがってそこでは, 行動を木構造に含まなかったとしても, 最低 fringe=5, すなわち $(n+2)^5$ もの状態表現を必要とする．USM に代表されるメモリーベース法は, 一般に, このように膨大な量のメモリーを必要とする．そのため現実的な手法とは言い難い．

ii. 確率的傾斜法との比較

次に確率的政策を学習する手法として確率的傾斜法 (SGA) [木村 96] に注目し比較を行う．SGA の学習率は 0.1, 割引率は 0.99 とした．また学習開始時に与える政策はランダム選択とした．

$n = 7, n = 14, n = 21$ としたときの PS-r*および SGA が学習終了時に獲得していた政策の質, ならびに学習に要した行動数を表 3 に示す．ここで政策の質は, 状態 X を始点としたときの報酬を得るまでに要する行動数で評価した．

表 3 図 11 の環境で PS-r*および SGA が学習終了時に獲得していた政策の質ならびに学習に要した行動数.

	n	政策の質		学習に要した行動数	
		平均	標準偏差	平均	標準偏差
PS-r*	7	24.2	0.218	2.38×10^4	4.10×10^4
	14	31.2	0.237	1.73×10^5	2.48×10^4
	21	38.2	0.237	2.19×10^5	6.77×10^4
SGA	7	26.3	6.83	2.21×10^3	2.91×10^3
	14	38.0	10.9	4.27×10^3	4.67×10^3
	21	50.8	9.47	4.42×10^3	5.47×10^3

PS-r*の学習の打ち切りには, $\alpha = 0.05$, $\beta = 0.10$ に設定した (6) 式を用いた．(6) 式の値は, π_1 と π_2 の差により大きく変化する．そこで, 本実験では, 検出すべき差の最大値を 0.05 に設定し (6) 式を用いた．この場合の (6) 式の最大値は 2095.09 となる．すべてのルールの選択回数がこの条件を上回った時点で学習を打ち切った．一方,SGA の学習は, 政策が収束した時点で打ち切った．

SGA は PS-r*に比べ学習は早いが, 得られる政策の質は劣ることがわかる．このことは n の値が大きくなるに連れ顕著になる．この主たる原因は, SGA では, 確率的政策が必要でない S_i においても, ある確率で無効ルールを選択してしまうためである．それに対し,PS-r*は状態 S_i では, 確実に無効ルールを排除することができる．そのために n が大きくなり状態 S_i 部分が長くなればなる程, SGA に対する PS-r*の優位性が増大する．

4·2 図 9a) の環境での評価

次に,PS-r および PS-r*のパラメータ r の影響を図 9a) に示す環境を用い調べる．

実験は乱数の種を変えて 100 回行った．初期値として与える r を変化させたときの PS-r*および PS-r が学習終了時に獲得していた政策の質, ならびに学習に要した行動数を表 4 に示す．政策の質は, 状態 0 を始点としたときの報酬を得るまで, すなわち, 状態 G に到達するまでに要する行動数で評価した．

PS-r*の設定は, 前節と同一である．PS-r の学習は, 新たな有効ルールが発見されなくなった時点で打ち切った．これは r の値には依存しないので, すべての r において PS-r が学習に要する行動数は同一である．また, ランダム選択が報酬を得るために要する行動数の平均は 5.49×10^2, 標準偏差は 5.19×10^2 である．

表 4 図 9a) の環境で PS-r*および PS-r が学習終了時に獲得していた政 策の質ならびに学習に要した行動数.

		政策の質		学習に要した行動数	
		平均	標準偏差	平均	標準偏差
PS-r*	0.8	36.7	6.12	5.24×10^5	3.15×10^4
	0.5	35.3	6.90	7.13×10^5	1.05×10^5
	0.1	36.7	5.30	6.02×10^6	8.51×10^5
PS-r	0.8	243	200	4.75×10^3	4.20×10^3
	0.5	101	68.4	4.75×10^3	4.20×10^3
	0.1	178	170	4.75×10^3	4.20×10^3

PS-r は r の値により得られる政策の質が大きく異な

る. r が 1 に近づくと状態 15 には適すが, 例えば, 状態 5 などでは, 下方の「袋小路」に遷移しやすくなる. 逆に,r が 0 に近づくと「袋小路」に陥るケースは減るが, 状態 15 と状態 14 の間を往復するケースが増える. それに対し,PS-r*では, 状態 5 などで「袋小路」に遷移することはなく, かつ, 状態 15 ではランダム選択をとることができる. そのため PS-r のように r の値によって学習の結果得られる政策の質が大きく変化することはない.

PS-r*では,r の値は検定時に使用される「ある特定の政策」に影響する. r が 0 に近付くと, 遷移しにくい状態が増えるので, 結果として, 検定の終了条件を満たすためには, 全体としてより多くの行動が必要となる. 一方, r が 1 に近付くと, 「ある特定の政策」とランダム選択との類似性が増すため, 検定の精度が低下する恐れがある. この環境でも,$r = 0.95$ とした場合には, 状態 15 に対し「不完全知覚なし」と判定する例が 100 回の実験中 29 回生じていた. 以上の結果より,PS-r*においては, 検定の精度および速度の観点から, r の初期値はあまり極端な値には設定すべきでないということが示唆される.

5. お わ り に

非マルコフ的環境の代表に POMDPs がある. 著者らは, 先に, ある特定の POMDPs 環境における, PS および RPM の合理性を保証する定理を導出した. しかしそこでは確率的政策が要求される一般の POMDPs 環境については議論されていない.

そこで本論文では, PS および RPM に確率的政策の考えを導入した手法である **PS-r** を提案し, 一般の POMDPs 環境下での挙動を解析した. その結果, PS-r が報酬を得るために要する行動数は, ランダム選択の最大 $(r\frac{(1+\frac{M-1}{r})^n}{M^n})$ 倍であることを示した. ここで n は不完全知覚により同一の状態とみなされている状態の個数の最大値, M は行動数である.

さらに確率的政策の利用を極力抑えることで, PS-r の挙動を改善する手法として **PS-r*** を提案した. PS-r*は,χ^2 検定を利用し不完全知覚状態を発見し, 不完全知覚状態のみでランダム選択を行う手法である. PS-r*は, 従来のメモリーベースよりも少ないメモリで, 従来の確率的政策を学習する手法では必ずしも明らかにされてはいないランダム選択からの改善を明示的に保証した手法である.

今後は,PS-r および PS-r*を報酬と罰が混在する環境 [宮崎 01] やマルチエージェント環境 [宮崎 99c] へ拡張することを行う予定である. また, 工学的応用も非常に重要であり, 有意義な適用例を早急に発見し, 提案手法の工学的な意味での有効性を主張したいと考えている.

謝　　辞

本稿をまとめるにあたり, 東京工業大学大学院総合理工学研究科の木村元助手から貴重なアドバイスをいただきました. ここに感謝の意を表します.

◇ 参 考 文 献 ◇

[Aberdeen 01] D. Aberdeen and J. Baxter: Scalable Internal-State Policy-Gradient Methods for POMDPs, Proc.of the 19th International Conference on Machine Learning, pp.3-10 (2002)

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

[Glickman 01] M. R. Glickman and K. Sycara: Evolutionary Search, Stochastic Policies with Memory, and Reinforcement Learning with Hidden state, Proc.of the 18th International Conference on Machine Learning, pp.194-201 (2001)

[Grefenstette 88] J. J. Grefenstette: Credit Assignment in Rule Discovery Systems Based on Genetic Algorithms, Machine Learning, Vol. 3, pp. 225-245 (1988)

[林 99] 林 朗, 末松 伸朗: POMDPs 研究に基づいたハイブリッド分類子システム, 人工知能学会誌, Vol. 14, No. 3, pp. 538-546 (1999)

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

[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)

[Konda 00] V. R. Konda and J. N. Tsitsiklis: Actor-Critic Algorithms, Advances in Neural Information Processing Systems 12, pp. 1008-1014 (2000)

[楠 98] 楠 正, 辻谷 将明, 松本 哲夫, 和田 武夫: 応用実験計画法, 日科技連 (1998)

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

[Loch 98] J. Loch and S. P. Singhf: Using eligibility traces to find the best memoryless policy in partially observable markov decision processes, Proc. of the 15th International Conference on Machine Learning, (1998)

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

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

[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 95] 宮崎和光, 山村雅幸, 小林重信: k-確実探査法：強化学習における環境同定のための行動選択戦略, 人工知能学会誌, Vol. 10, No. 3, pp. 454-463 (1995)

[宮崎 98] 宮崎和光, 小林重信: POMDPs における合理的政策の逐次改善アルゴリズムの提案，第 25 回知能システムシンポジウム資料，pp.87-92 (1998)

[宮崎 99a] 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 99b] 宮崎和光, 木村元, 小林重信: Profit Sharing に基づく強化学習の理論と応用, 人工知能学会誌, Vol. 14, No. 5, pp. 800-807 (1999)

[宮崎 99c] 宮崎和光, 荒井幸代, 小林重信: Profit Sharing を用いたマルチエージェント強化学習における報酬配分の理論的考察, 人工知能学会誌, Vol. 14, No. 6, pp. 1156-1164 (1999)

[Miyazaki 99] K. Miyazaki and S. Kobayashi: Proposal for

an Algorithm to Improve a Rational Policy in POMDPs, IEEE International Conference on Systems, Man and Cybernetics, Vol. V. pp. 492-497 (1999)
[宮崎 01] 宮崎和光, 坪井創吾, 小林重信: 罰を回避する合理的政策の学習, 人工知能学会論文誌, Vol. 16, No. 2, pp. 185-192 (2001)
[Singh 94] S. P. Singh: Reinforcement Learning Algorithms for Average-Payoff Markovian Decision Processes, Proc.of the 12th National Conference on Artificial Intelligence, Vol. 1, pp. 700-705 (1994)
[末松 98] 末松 伸朗, 林 朗, 李 仕剛: 部分観測環境での強化学習へのモデルベースアプローチ：可変長記憶モデルのベイズ学習, 人工知能学会誌, Vol. 13, No. 3, pp. 404-418 (1998)
[Sutton 98] R. S. Sutton and A. Barto: Reinforcement Learning: An Introduction, A Bradford Book, The MIT Press (1998)
[Sutton 00] R. S. Sutton, D. McAllester, S. P. Singh and Y. Mansour: Policy Gradient Methods for Reinforcement Learning with Function Approximation, Advances in Neural Information Processing Systems 12 (NIPS12), pp. 1057-1063 (2000)
[Watkins 92] C. J. H. Watkins and P. Dayan: Technical note: Q-learning, Machine Learning Vol. 8, pp. 55-68 (1992)
[Whitehead 95] S. D. Whitehead and L. J. Lin: Reinforcement learning of non-Markov decision processes, Artificial Intelligence 73, pp271-306 (1995)
[Wiering 96] M. Wiering and J. Schmidhuber: Solving POMDPs with Levin Search and EIRA, Proc. of the 13th International Conference on Machine Learning, pp.534-542 (1996)
[Williams 92] R. J. Williams: Simple Statistical Gradient Following Algorithms for Connectionist Reinforcement Learning, Machine Learning, Vol. 8, pp. 229-256 (1992)

〔担当委員：新谷虎松〕

2002 年 10 月 22 日 受理

♢ 付　　録 ♢

A. 定理 3.1 の証明

一般にタイプ 2 の混同が存在しなければ, 無効ルールを選ぶ必要はない. ランダム選択ではすべてのルールが等確率で選択されるので, 結果的に無効ルールと有効ルールの選択確率は等しくなる. 一方, PS-r では無効ルールの選択確率は有効ルールよりも低い. したがって, タイプ 2 の混同が存在しなければ, PS-r が報酬を得るために要する行動数がランダム選択の行動数よりも多くなることはない.

B. 補題 3.1 の証明

すべてのルールが有効ルールのときはランダム選択と一致するので, 以下では, 無効ルールが存在する場合のみを考える.
(i) $M = 2$ の場合
構造ωは, 最も報酬に近い状態 (状態 B) 以外では, 無効ルールを選ぶと必ず報酬に近づき, 有効ルールを選ぶと必ず報酬から最も遠い状態 (状態 A) に遷移する構造である. したがって, 構造ωは,$M = 2$ のすべての構造の中で, 無効ルールを選べば選ぶほど最も報酬に近づき, 有効ルールを選べば選ぶほど最も報酬から遠ざかる構造である. すなわち構造ωでは, 有効ルールを選べば選ぶほど, 報酬獲得までの行動数は多くなる. 構造ωよりこの効果が弱い, すなわち有効ルールを選択したときの報酬獲得までの行動数の悪化が, 構造ωより少なくなる構造では, よりランダム選択との差が縮まる. したがって, $M = 2$ のときは, 構造ωが最も PS-r とランダム選択との差が開く構造である.
(ii) $M > 2$ の場合
$M = 2$ の構造に第 3 のルールを加えたとき, そのルールの選択確率がゼロでない限り, 報酬獲得までの行動数は $M = 2$ の場合より増加する.

加えたルールが無効ルールの場合, ランダム選択ではその第 3 のルールの選択確率は $\frac{1}{M}$ となるが, PS-r では $\frac{1}{M}$ より小さい確率で選択される. したがって, 同じ構造で比較する限り, そのような (第 3 の) 無効ルールが増えれば増えるほどランダム選択と PS-r との差は縮まる.

一方, 第 3 のルールとして有効ルールを加えた場合は, 行動 b を出力するルールと同一のルールが増えれば増えるほど, PS-r とランダム選択との差が開く. したがって, 構造ωで行動 b を出力するルールと同一のルールが最大, すなわち $M - 1$ 個存在する場合が最もランダム選択との差が開く場合である.

以上より, 構造ωが, 最も PS-r とランダム選択との差が開く構造となる.

C. 補題 3.2 の証明

構造ωで, 報酬獲得までのステップ数 V_a は

$$V_a = \frac{1}{s(1-s)^{n-1}}, \tag{C.1}$$

となる. ここで s は, 構造ωで $M - 1$ 個存在する行動 b を出力するルールの選択確率であり, ランダム選択の場合 $s = \frac{M-1}{M}$, PS-r の場合 $s = \frac{M-1}{(M-1)+r}$ となる. これらの s を (C.1) 式に代入し, 比率を計算すると, 構造ωで PS-r が報酬を得るまでに要する行動数は, ランダム選択の最大 $(r\frac{(1+\frac{M-1}{r})^n}{M^n})$ 倍となる.

著　者　紹　介

宮崎　和光(正会員)

1991 年明治大学工学部精密工学科卒業. 1996 年東京工業大学大学院総合理工学研究科知能科学専攻博士後期課程修了. 博士 (工学). 同年 4 月, 同大学大学院総合理工学研究科助手. 1998 年 4 月, 同大学大学院総合理工学研究科リサーチアソシエイト. 1999 年 10 月, 学位授与機構審査研究部助教授. 2000 年 4 月, 大学評価・学位授与機構学位審査研究部助教授. 現在に至る. 人工知能, 特に強化学習に関する研究に従事. 計測自動制御学会, 日本機械学会, 日本高等教育学会各会員.

小林　重信(正会員)

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