エピソード記憶編集による迷路課題の学習アルゴリズムの提案

青田佳士[†] 山口陽子[††]

学習器が新奇環境での規則を新たに獲得する際には，古い規則との干渉をいかに避けるかが難しい問題である．そのため Profit sharing といったナビゲーションによく利用される学習アルゴリズムでは一般に，Alternation maze task のようにゴールの位置が交互に変わるような迷路課題に直接適用することができず，行動の開始と終了を自律的に決定するメカニズムが必要とされる．本研究では，成功や失敗といった経験に基づいたエピソード記憶編集による新しい学習アルゴリズムを提案する．このアルゴリズムを利用して経験から隠れたルールを見い出すことにより，ゴールの位置が複雑に変わるような課題も学習可能となることを示す．また，課題の変化が積み重ねとして，かえって学習にプラスの効果をもたらす場合についても考察する．

An Algorithm for Solving Maze Tasks with Episodic Memory Integration

Yoshito Aota[†] and Yoko Yamaguchi[††]

Learning of a new law in a novel environment is known to be a difficult problem because of interference of the old law. In an alternation maze task where goal position changes alternatively in each trial, popular algorithms such as Profit sharing are not directly available without some additional algorithm to avoid interference of past experiences. An autonomous mechanism which decides start and end of behaviors is needed. In this research, we propose a new learning algorithm using episodic memory integration based on successful and failed experiences. We demonstrate that this algorithm enables learning of subsequently changing goals by finding a hidden rule from previous experiences. It is also discussed that change of tasks gives positive learning effect rather than negative effect, as accumulated experience.

1. はじめに

機械学習では，不確定な環境として一般に部分観測マルコフ決定過程（POMDPs）を仮定した問題が取り組まれる[1)]．POMDPs 下では，エージェントは実際には異なる環境の状態を同一の知覚入力と見なすことがある．そのためエージェントの知覚センサや移動先などに確率的な要因を設定したり，空間上のある位置と別の位置の知覚入力を同一にしたりすることで POMDPs 下の課題として取り組まれるのが一般的である．

一方で，エージェントの位置と知覚入力を 1 対 1 対応にして位置における不確定性をなくしたとしても，エージェントがその位置にいるタイミングなどの状況によって異なる行動を選択しなければならないような課題も考えられる．

† 横浜国立大学大学院国際社会科学研究科
International Graduate School of Social Science, Yokohama National University
†† 理化学研究所脳科学総合研究センター
Brain Science Institute, RIKEN

そのような課題として Alternation maze task[2)] が知られている．これはラットに迷路を解かせる課題の 1 つで，**図 1** のような迷路においてゴールの位置がラットの行動に応じて左右交互に変わっていく課題である．ラットはこの種の迷路課題に対して状況依存的[3)]に知能を働かせ，比較的容易にゴールの位置を予測できることが分かっている．この課題の特徴は，ゴールの位置が変化する以外の環境変化を仮定しない点にある．ラットはそのため，知覚入力が同じでも自分がどのタイミングでその位置にいるかに応じて異なる行動選択をとらなければ適切にゴールにたどり着くことができない．

Alternation maze task も POMDPs に含まれる課題であるが，隠れ状態はゴールの出現タイミングである．そのためエージェントがこの課題を解くには，ゴールの位置がどのように変化するかを学習しなければならない．それにはゴールに着いた時点で時間ステップ t をリセットするというように一連の行動の区切り方を固定することはせず，時間軸上のどこで行動を区切るかは不確定な状態から始める必要がある．

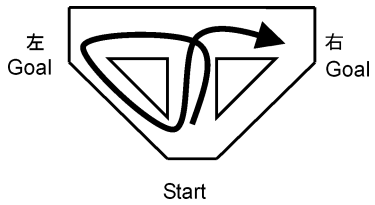

図 1 本研究で用いる Alternation maze task の軌道の例．
左右の道端を交互に訪ねることでゴールで報酬を受ける
Fig. 1 Example of Alternation maze task.
Goal position changes alternatively in each trial.

POMDPs に対するアプローチとしては従来，Profit sharing に代表される経験強化型の強化学習（たとえば文献 4)）やメモリベース法（たとえば文献 5)），あるいは確率的政策を学習する手法（たとえば文献 6)）がとられてきた．これらの手法を Alternation maze task にあてはめるには「はじめ」に左のゴールを選んだ後に右のゴールを選んで「終了」といった形で，課題の始まりと終わりを明確に与える必要がある．そのため，同じ POMDPs に分類されても，とくに時系列の「始まり」と「終わり」を自律的に決めなければならない課題の場合，従来手法では解決に困難をともなうと考えられる．

実際 Profit sharing ではエピソードとして，メモリベース法では履歴の木構造として個々の状況を時間的に区切っているが，その区切りの仕方はゴールで時間ステップ t をリセットする形である．たとえば Profit sharing において過去のエピソードを知覚入力に加える際に，図 1 の迷路でゴールが左右交互ではなく「左左右」と変わる課題には過去 2 つ分のエピソードが必要など，過去のいつまで遡ればよいのかをあらかじめ決めることができない．メモリベース法では，行動の開始時点を決めて履歴の木構造をたどる必要があるが，これも開始時点をあらかじめ決めることができない．また，確率的政策では，エージェントの観測–行動のペア（X, a）に内部変数パラメータ W を加えた確率 π(a, W, X) に従って行動 a が実行されるとした場合，内部変数パラメータ W を適切に学習させることで POMDPs の環境に広く対応できうることが知られている[6]．しかし，内部変数パラメータの適切な数や学習のさせ方は一般に既知でなく，とくに時間ステップ t の柔軟な制御はあまり考慮されてこなかった．

したがって，一連の行動の開始と終了を自律的に決め，学習によって状況を時間的にも積極的に分離する仕組みが必要である．

本研究では，個々のゴールに対する最適性よりもゴールの継続的獲得を目指す．ゴールの位置が変化する場合，Profit sharing のようにエピソードを重ね合わせることで最適解を求めるよりも，ゴールの経験を別々に蓄積し組み合わせるようにできることが有効と考えられる．そのためここでは，Profit sharing と同じエピソードの枠組みを利用しつつ，さらにエピソードを別個に貯蔵し状況に応じて想起・実行・編集することで次の行動計画を立てる仕組みを提案する．これにより，POMDPs 下でかつ行動の開始と終了を自律的に決定しなければゴールの位置変化の規則を獲得できないような課題を学習できたことを述べる．

2 章では学習アルゴリズムについて，およびエピソード編集のダイナミクス例を述べる．3 章では実験に用いられた迷路課題を説明し，エージェントを適用した結果や Profit sharing との比較実験結果を示す．4 章では考察として，学習の積み重ね効果や非決定的な状態遷移が存在する場合，および解の最適性について述べ，5 章でまとめを行う．

2. 学習アルゴリズム

2.1 課題設定とエージェント内の情報表現

本論文でエージェントが解く迷路課題では，ゴールの位置がある規則で変化する．エージェントはこの隠れた規則を学習しなければならない．観測状態はエージェントの位置と報酬の有無であり，隠れ状態は各位置で報酬が観測されるタイミングである．

本研究では POMDPs 下でもとくにゴールの位置変化の規則をいかに学習するかを考えたいので，非決定的な状態遷移については取り扱わなかった．非決定的環境におけるエージェントの振舞いについては，4 章の考察で議論する．

環境は，マス目に分割された枝分かれを含む通路とし，特定の場所（ゴール）に一定の規則で報酬がおかれる．エージェントの行動は時間，空間とも離散値で表される．単位時間を 1 ステップとして，1 マスを 1 ステップで移動する．行動は各ステップにおけるマス目固有の知覚入力と，移動の方向とで記述される．

移動の結果，ゴールに到着して報酬を得る場合を成功と呼ぶ．行動計画の中に報酬の予測が含まれていたにもかかわらず，実際にそのマス目で報酬が得られなかった場合を失敗と呼ぶ．一連の行動を成功か失敗の直後で区切ったものをエピソードと呼び，さらにエピソードの終わりが成功で区切られたもの（成功を含むエピソード）を成功エピソード，失敗で区切られたものを失敗エピソードと呼ぶ．たとえば各マス目を数字で表したとして，「1 → 2 → 3 → 4 → 6 → 7 → 1 → 2 → 3 → 4 → 5 + 報酬」で 1 つの成功エピソー

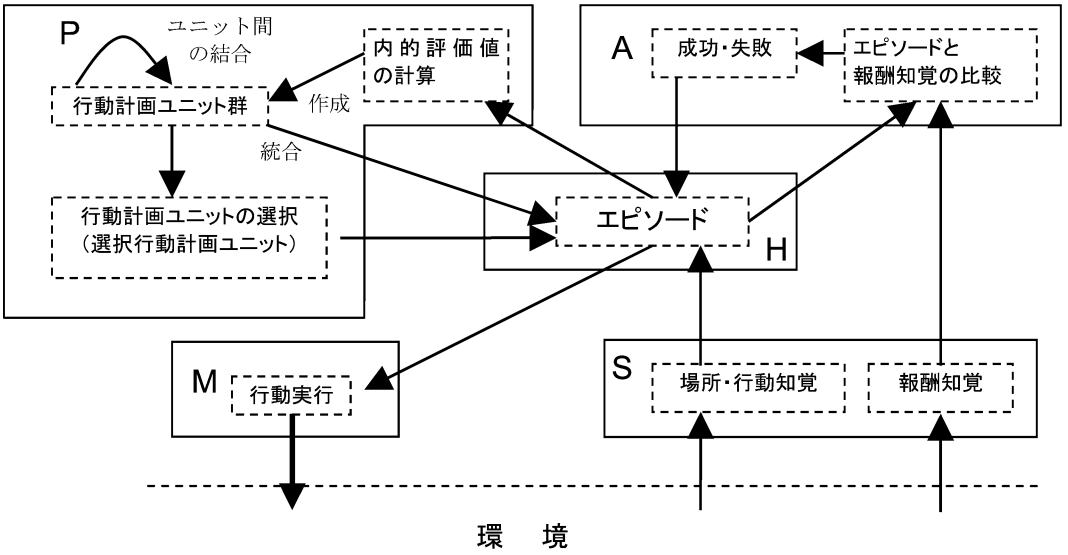

図 2 ネットワーク構造.
S は知覚入力，H はエピソードの形成・貯蔵，A は成功・失敗の判断，P は行動計画ユニットの作成・行動計画とエピソードの編集，M はエピソードを通じて選択行動計画ユニットの実行を行う
Fig. 2 Network Structure.
S: sensory input part, H: store of episods, A: decision of success and fail, P: episodic memory integration, and M: motion part.

ドといった具合である．つまり Profit sharing のエピソードの定義に失敗も含めたものである．

1 つまたは複数のエピソードから構成される時系列を行動計画ユニットと呼ぶ．行動計画ユニットは行動の選択肢として利用されるが，行動計画ユニットとして構成されない単に複数のエピソードが続いたものを連続エピソードと呼ぶ．そのため以降では，1 つのエピソードと複数の連続したエピソードの両方を指して「(連続) エピソード」と呼ぶことがある．さらに各行動計画ユニットはその中に含まれる成功と失敗の回数，および経過ステップ数より決定される内的評価値を 1 つの値として持つ (2.4 節参照)．行動計画ユニットがどのように作成されるかは，次節で述べる．初期状態 (実験の開始時)，あるいは成功の直後を開始として次の成功を終了とする時系列を 1 試行と呼ぶ．したがって 1 試行終了までに何度か失敗を経験することもある．

2.2 ネットワーク構造の概要

提案モデルの全体をネットワーク構造として表すと**図 2** のようになる．図 2 で S 層は知覚入力層で，エージェントの現在位置や移動方向，報酬の有無を H 層および A 層に送る．A 層は成功・失敗を判断する層で，前節で定義したように報酬を得た場合を成功，H 層から入力される報酬の予測がはずれた場合を失敗として H 層に出力する．H 層では A 層の成功・失敗の情報に基づいてエピソードが形成・貯蔵される．1 つのエピソードは 1 つの成功か失敗を含む．

P 層は，H 層で貯蔵されたエピソードに基づいて行動計画ユニットを作成・編集し，次の行動計画を立てる．エージェントの行動計画は，行動計画ユニット単位で決まる．行動計画ユニットは，以下の場合でかつ該当する (連続) エピソードが初めて経験したものであるとき，それを 1 つの行動計画ユニットとして P 層に保存する．

＜行動計画ユニット作成対象のエピソード＞

- **報酬を得たとき**の下記のエピソード
1. 探索行動で報酬を得たエピソード
2. 選択行動計画ユニットがコードしている報酬の獲得タイミングと異なるタイミングで報酬を得た場合で，行動開始から最後の報酬までの (連続) 成功エピソード
- **直前が失敗のとき**の下記のエピソード
3. 行動開始から直前の失敗までの ((連続) 成功エピソード +) 失敗エピソード
4. その前の失敗から直前の失敗までの連続成功エピソード

これにより行動計画ユニットは複数のエピソードとして随時編集され，開始と終了のタイミングが様々な

① 開始時の行動決定

1) 前の行動計画ユニットと結合のある行動計画ユニットを，なければ，現在の知覚入力と結合のある行動計画ユニットを想起する
2) 想起した行動計画ユニットで内的評価値の高い行動計画ユニットを選択行動計画ユニットにする
3) 選択過程に失敗・報酬に応じたランダムネスを仮定し，適当な頻度で探索行動を行う

② 選択行動計画ユニットの実行

選択行動計画ユニットの中に，

1) 成功をコードしているエピソード（成功エピソード）を含むとき，成功エピソードを実行．失敗をコードしているエピソード（失敗エピソード）もあれば，その時点で ① へ
2) 失敗エピソードのみを含むとき，探索行動を行う

③ 選択行動計画ユニット実行後の記憶の編集

以下の場合，該当する（連続）エピソードが初めてのとき，それを 1 つの行動計画ユニットとして保存する

1) 探索行動により報酬を得た場合，そのエピソード
2) 選択行動計画ユニットがコードしている報酬の獲得タイミングと異なるタイミングで報酬を得た場合で，行動開始から最後の報酬までの（連続）成功エピソード
3) 直前が失敗のとき，その行動開始から直前の失敗までの（(連続）成功エピソード＋）失敗エピソード
4) 直前が失敗のとき，その前の失敗から直前の失敗までの連続成功エピソード

また，新しい行動計画ユニット作成の有無にかかわらず，上記 2)，3) のときは開始時の選択行動計画ユニットと 2 つ前の行動計画ユニットとの結合を抑制，1)，2) のときは手前 2 つの行動計画ユニットを正の結合，4) ではその行動計画ユニットの直前の，失敗エピソードを含む行動計画ユニットと 4) を正の結合にする

図 3 学習アルゴリズム
Fig. 3 Learning algorithm.

行動計画ユニットが作成される．こうして行動の開始と終了は自律的に決定されることになる．行動計画ユニットはエピソード単位で行動を開始・終了するため，ゴールの位置の変化に対応しやすい．

M 層は実行層で，選択行動計画ユニットに基づいた行動をとる．選択行動計画ユニット中の成功エピソードと同じ行動をとり，失敗エピソードでは探索行動を実行するか次の行動計画に移る．

2.3 行動計画のアルゴリズム

図 3，**図 4** に，より具体的なアルゴリズムを示した．

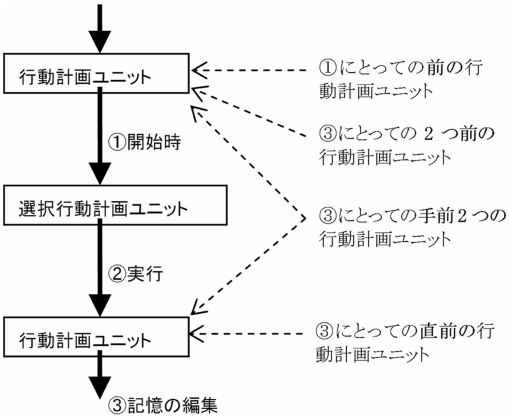

図 4 行動計画ユニットの順序関係．
学習アルゴリズム（図 3）の各段階（①–③）の説明で出る「行動計画ユニット」がどれを指すかを説明
Fig. 4 Relationship between ①–③ of Fig. 3 and Action plan units.

基本方略としては，大きく以下のような流れで学習を進める．

① 開始時の行動決定

状況（前回の行動計画ユニットと知覚入力）に応じて行動計画ユニットを想起し，より内的評価値の高い行動計画ユニットを選択．これを選択行動計画ユニットとする．

② 選択行動計画ユニットの実行

(1) 選択行動計画ユニットの中に，成功エピソードを含むとき，それを実行し，失敗エピソードもあれば，その時点で行動終了として ① に戻る．
(2) 選択行動計画ユニットが失敗エピソードのみから構成されるとき，探索行動を行う．

③ 選択行動計画ユニット実行後の記憶の編集

選択行動計画ユニットがコードするエピソードと実際に経験したエピソードを比較し，報酬と経験の有無に応じて新しい行動計画ユニットを作成，さらに行動計画ユニット間の統合や順序関係の学習を行う．

なお，数式表現によるアルゴリズムの詳細については付録を参照されたい．

2.4 内的評価関数

行動計画ユニットの作成時に決められる内的評価値 w^{ER} は，下記の内的評価関数に従う．

$$w^{ER} = \begin{cases} 4(b^2/t), & \text{成功時} \\ 4(b^2/t_{\text{success}}) + (2/t_{\text{fail}}), & \text{失敗時} \end{cases}$$

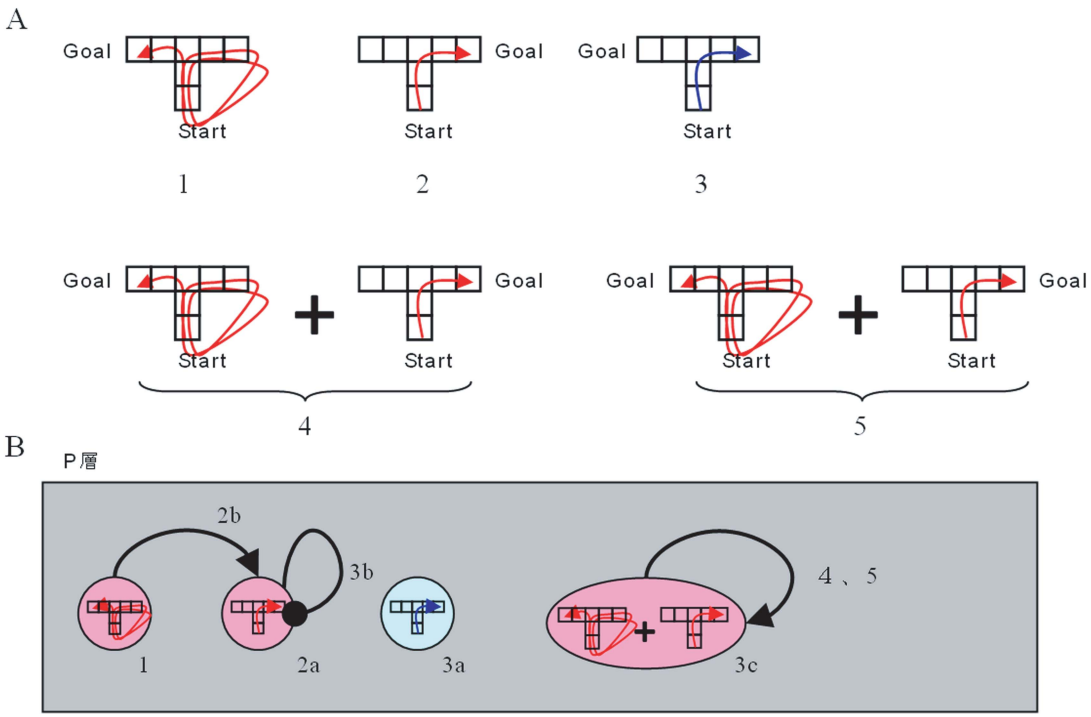

図 5 エージェントが Alternation maze task に取り組んだときの例.
A：実際にエージェントがとった行動．赤い矢印は成功時の行動，青い矢印は失敗時の行動を表す.
B：P 層で作成された行動計画ユニットとその結合で “⟶” は正の結合を表し，“—●” は負の結合を表す．番号は A と対応し，新しい行動系列を新しい行動計画ユニットとして P 層に作成し，図 3 のアルゴリズムに従ってユニット間に結合を作る．本文 2.5 節参照

Fig. 5 Example of the agent's learning process.
A: actual behaviour of the agent. Red arrows show successful behaviour, and a blue arrow shows failure behaviour.
B: connections between Action plan units which were made in the P layer. The arrow “⟶” is excite connection, and the arrow “—●” is inhibit connection. Refer the section 2.5.

ここで b はその行動計画ユニットがコードする報酬の数で，t は総ステップ数，t_{success} は開始から最後に報酬を得たときまでのステップ数，t_{fail} は最後に報酬を得たときから失敗したときまでのステップ数である．そのため，$t = t_{\mathrm{success}} + t_{\mathrm{fail}}$ を満たす.

w^{ER} は，報酬の数が多いと高い値をとり，ステップ数が大きいと低い値となる.

2.5 P 層の行動計画ユニットのダイナミクス

図 2 のようなネットワーク構造をとるエージェントが Alternation maze task を解く場合の，P 層の行動計画ユニットのダイナミクスを**図 5** に示す．図 5 の A はエージェントが各試行で実際に経験したエピソードを表し，図 5 の B では P 層に作成された各行動計画ユニットを円もしくは楕円で表した．図 5 の B で矢印は行動計画ユニット間の正の結合を示し．終点が黒丸の線は負の結合を示す.

まず，エージェントが初めて課題に取り組んで図 5 の A の 1 のようなエピソードを経験したとき，P 層では図 5 の B の 1 のように 1 つの成功エピソードからなる新しい行動計画ユニットとして貯蔵される．この時点でエージェントが持つ行動計画ユニットはこの 1 つのみなので，ランダムネスを考慮しなければ，次の試行でも前の行動計画ユニットと同じ行動を実行しようとする.

ところが Alternation maze task ではゴールの位置が先ほどの左から右に変わっているので，エージェントが経験した実際のエピソードは図 5 の A の 2（= 図 5 の B の 2a）のようになる．図 5 の B の 1 と 2a の行動計画ユニットでは 2a のほうが開始から報酬までの総ステップ数が少ない，すなわち内的評価値が高いので，エージェントは再び図 5 の A の 2 の行動系列を実行する．しかしゴールの位置が右から左に変わっ

ているので，エージェントは図 5 の A の 3 のような失敗エピソードを経験する．そのとき P 層では，この失敗エピソードからなる行動計画ユニットと，以前に経験した 2 つの成功エピソードからなる行動計画ユニットが作成される（図 5 の B の 3a，3c）．そうして，この 2 つの成功エピソードをコードする行動計画ユニットが内的評価値の最も高い行動計画ユニットとなり，次の試行からはそれが選択行動計画ユニットとして選ばれる．

この例で分かるように，エージェントは次のゴールの位置を予測できても，その経路は必ずしも最適ではない．本研究では継続的なゴールの経験をまずは優先するが，解の質については事前学習によりある程度改善されうることを 3 章の結果で見，4 章で考察する．

3. 実験課題とシミュレーション結果

3.1 Profit sharing との比較実験

3.1 節では，T 字型迷路において Profit sharing との比較実験を行い，本アルゴリズムの有効性を検証する．

（1） **T 字型迷路：予備段階あり**

まず予備段階 1 として，ゴールの位置が**図 6** の A に固定された場合を 200 試行，次に予備段階 2 として B に固定された場合を 200 試行学習する．そのうえで本課題を行う．本課題のゴールの位置は，エージェントがゴールにたどり着くたびに A → A → B → A → A → B → ... と変わる．つまり AAB の 3 試行からなる周期を繰り返す．この課題を適切に学習するには，空間的には同じ A の行動系列を時間的に異なるエピソードとして区別する必要がある．また，予備段階が本課題の学習にどのように影響するかも確認する．

なお，本論文ではゴールの位置の変化があるパターン（たとえば AAB の 3 試行で 1 つのパターン）で繰り返される課題を対象とする．学習の収束まで何回そのパターンが繰り返されたかなどを見たいので，主として，1 つのパターンを 1 周期とした「周期」を単位に取り扱う．

POMDPs 下の課題は，最適解を得るのが一般に困難である．ただし本研究で扱う迷路課題では決定的な合理的政策が存在するので，分岐点でランダム選択を実行すれば確実にゴールに至ることができる．T 字型迷路をランダム選択によって解く場合，確率 0.5 で 4 ステップ，確率 $(0.5)^2$ で $4+5$ ステップ，...，確率 $(0.5)^n$ で $4+5\times(n-1)$ ステップかかるので，ゴールに至るまで平均で 9 ステップかかる．そのためステップ数が 9 より小さくなるかどうかが学習アルゴリズムの有効性の指標になる．

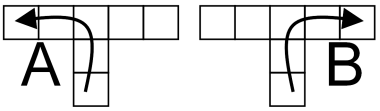

図 6 T 字型迷路で用いられる 2 つの試行 A と B
Fig. 6 Trials A and B of T maze task.

（2） **T 字型迷路：本課題のみ**

(1) で予備段階を経ず，本課題のみの場合．

比較のために Profit sharing を次のようにあわせて実行した．Profit sharing としては，仮にゴールが固定されていた場合に合理性定理[7] を満たすよう，強化関数 f として以下のような等比減少関数を用いた[8]．

$$f(t, R_T, T) = \gamma^{T-t-1} R_T$$

ここで T はエピソードが終了した時刻（ステップ数）であり，γ は割引率パラメータで 0.25 とし，R_T は時刻 T で得られる報酬で 1 とした．また，行動選択に重み付きルーレット選択を用いた[8]．

$$\Pr(s,a) = W(s,a) \Big/ \sum\nolimits_{a' \in A(s)} W(s,a')$$

$\Pr(s,a)$ は状態 s で行動 a を選択する確率，$A(s)$ は状態 s で実行可能な行動の集合を表す．W の初期値は 0.1 とした．

以上において，本アルゴリズムと Profit sharing の各 1000 個体が上記 (1)(2) の課題に取り組んだ．

3.2 T 字型迷路のシミュレーション結果

図 7 に各周期あたりのゴールまでに要した平均ステップ数を示す．はじめの 200 周期まではゴールの位置が図 6 の A に固定されており，その場合の学習効率は Profit sharing の方が良い結果を示している．しかし Profit sharing はゴールの位置が B に移った時点で大きな干渉を示し，また AAB と周期的なルール（本課題）に変わるとランダム選択とほぼ同じ結果となった．「本課題のみ」に Profit sharing を適用した場合はランダム選択よりも悪い結果であった．

一方で本アルゴリズムでは，「予備段階あり」と「本課題のみ」の両方でランダム選択よりも良い結果を示した．また，「予備段階あり」と「本課題のみ」との間では「予備段階あり」の方が少ないステップでゴールに到達でき，予備段階の経験が有効に働いたことが分かった．さらに，ゴールの位置が 200 周期目以降で A から B に移ったときの干渉の程度は Profit sharing よりもずっと小さかった．

図 8 には本課題における本アルゴリズムの学習曲線を示した．「予備段階あり」では若干の干渉が見ら

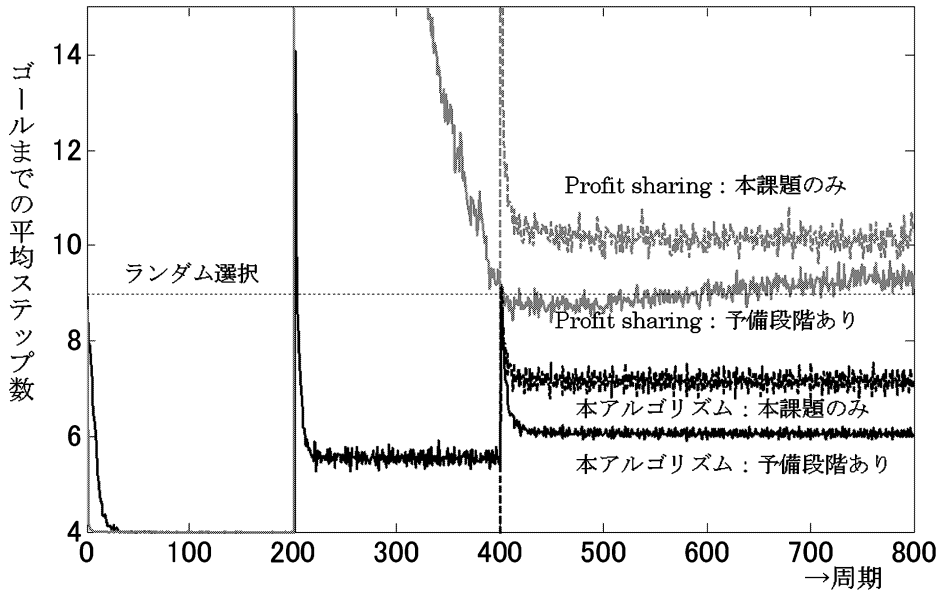

図 7 T 字型迷路による比較実験結果.
ゴールの位置は 200 周期までは図 6 の A，400 周期までは B で固定，以降は AAB の順に変わる．本モデルは学習の干渉をそれほど示さないのに対し，Profit sharing では 200 周期を過ぎた時点で大きな干渉を示し，また，400 周期以降はランダムに近い結果となった

Fig. 7 Result of the comparative experiment in the T maze tasks.
Goal position is fixed at A until 200 trials, and B until 400 trials. It is changed in the order AAB after 400 trials. Profit sharing showed large interference after 200 trials and near the result of random behaviour after 400 trials, though the proposed model didn't show such large interference.

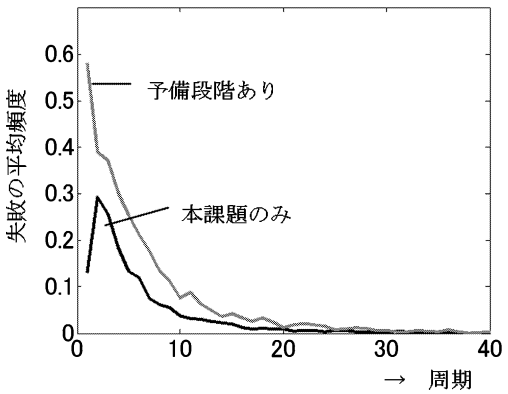

図 8 T 字型迷路の学習曲線：予備段階ありは学習に余分に時間がかかる

Fig. 8 Learning curve of the T maze task: Agents showed a little of interference after the pre-learning.

れたが，両条件とも 30 周期までにほぼすべての個体がゴールの位置を正確に予測できるようになった．

以上より，本アルゴリズムは Profit sharing と比べてゴールの位置の変化に頑健であり，次のゴールの位置を間違いなく予測できること，そして過去の経験は新規の課題に対して学習をやや遅らせはするが，解の質には有効に作用しうることが確認できた．

3.3 問題解決のプロセス

本アルゴリズムにおける問題解決の原理を見るために，「予備段階あり」と「本課題のみ」の場合の典型的な例について調べる．**図 9** は，各ステップにおいて選択された，もしくは形成された行動計画ユニットを指す．A は「予備段階あり」，B は「本課題のみ」の例である．横軸は周期ではなく，課題開始からの累計ステップ数を示した．

この例において最終的に形成された行動計画ユニットは，「予備段階あり」の場合が 11 番目の行動計画ユニットで，「本課題のみ」の場合は 13 番目の行動計画ユニットであった（図 9）．「予備段階あり」における 11 番目の行動計画ユニットは AAB に対応した 3 つのエピソードから構成されていた（**図 10**）．しかし「本課題のみ」における 13 番目の行動計画ユニットでは，10 以上のエピソードや他の行動計画ユニットから構成された長い周期として解を出していた．

「予備段階あり」の例の場合，AABAABAAB... からなる本課題のはじめにおいて，予備段階の学習において形成された A や B の成功エピソード単体からな

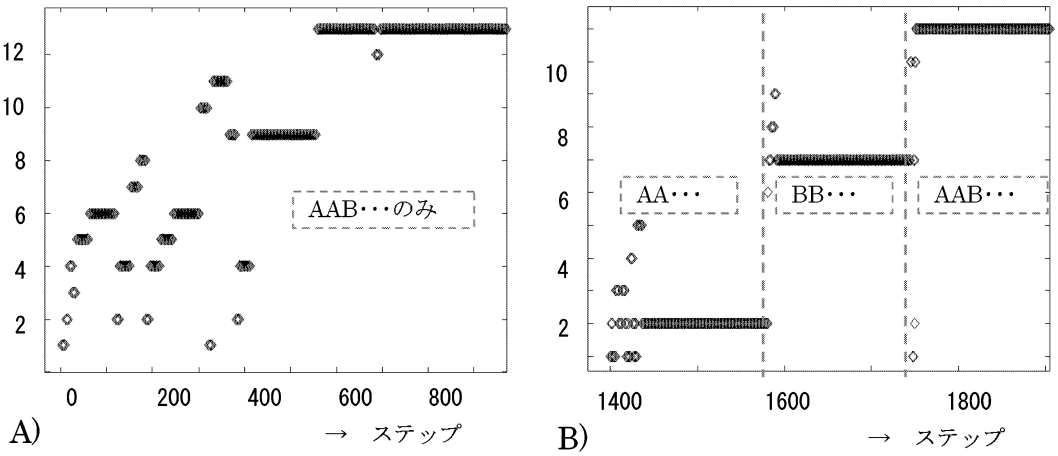

図 9 T 字型迷路課題 (1)(2) における P 層のダイナミクス例.
縦軸は行動計画ユニットの番号を表す．A）が「本課題のみ」の例で，B）が「予備段階あり」の例．予備段階ありでは B の 11 番目の行動計画ユニットは本課題の周期に対応していたが（図 10 も参照），A の 13 番目の行動計画ユニットは本課題の周期に対応せず，ABAABAABA といった長い系列となっていた

Fig. 9 Example of the P layer's behaviour in the T maze tasks (1) and (2).
The vertical axis shows number of Action plan unit. A) is an example of learning process without pre-learning. B is the process with pre-learning. The 11th unit of B) corresponded the period of goal change AAB, but the 13th unit of A) showed long period like ABAABAABA.

る行動計画ユニットが優先的に選択されたために，本課題と同じ周期の行動計画ユニットの形成がなされていた（図 10）．この例では，図 3 の ③-4 の学習則により 1，2，7 番目の行動計画ユニットが AAB に対応する 1-7-2 の行動計画ユニット 11 として形成された．したがって，a) 過去に経験した系列が現在学習したい系列に部分的にでも反映されていること，b) 連続した成功エピソードを 1 つの行動計画ユニットとして統合すること（図 3 の ③-4 の学習則）の 2 つが学習に重要と考えられる．

3.4 十字型迷路による学習の収束性の検証

本研究では，エージェントは後方への移動や壁への衝突はないとし，前方か左右の 3 方向への移動しかできないものとした．後方への移動や壁への衝突は，場合分けの数が増えることを意味する．つまり探索行動が増え，学習の収束までより多くのステップ数を要する．

そこでこの節では分岐の数を 1 つ増やした十字型迷路を用い，さらにルールを複雑にした場合を考える．これによりエージェントの探索行動の機会を増やし，本アルゴリズムの学習の収束性を確認する．用いた迷路課題は次のとおり．

(1) 十字型迷路：予備段階あり

予備段階としてゴールの位置が A に固定されている場合を 200 試行学習後，B に固定されている場合を 200 試行，C の場合を 200 試行学習する（**図 11**）．次に本課題に移る．本課題でのゴールの位置はエージェントがゴールにたどり着くたびに図 11 のように 20 試行を 1 つの周期として順に変わる．1 周期が 20 試行と長い系列であるだけでなく分岐点におけるエージェントの選択肢が 3 つに増えて難しい課題であるため，学習の収束能力を確認できる．

(2) 十字型迷路：本課題のみ

(1) で予備段階を経ず，本課題のみの場合．

以上において各 400 個体のエージェントが課題に取り組む．

3.5 十字型迷路のシミュレーション結果

図 12 は学習曲線を，**図 13** はゴールまでの平均ステップ数を示す．黒線とグレーの線はそれぞれ「本課題のみ」と「予備段階あり」を表す．

図 12 によると，「本課題のみ」と「予備段階あり」の両条件とも最終的には誤ったエピソードを選択しなくなったことが分かった．1 周期が 20 試行と長い系列ではあるが，ゴールを予測するエピソードを正しく選択し，失敗を含まない行動計画ユニットを形成できるようになったことを意味する．また，図 13 によるとランダム選択よりも少ないステップ数でゴールに到達しており，解の質もより良いものであることが示された．図 13 では「予備段階あり」の方が「本課題の

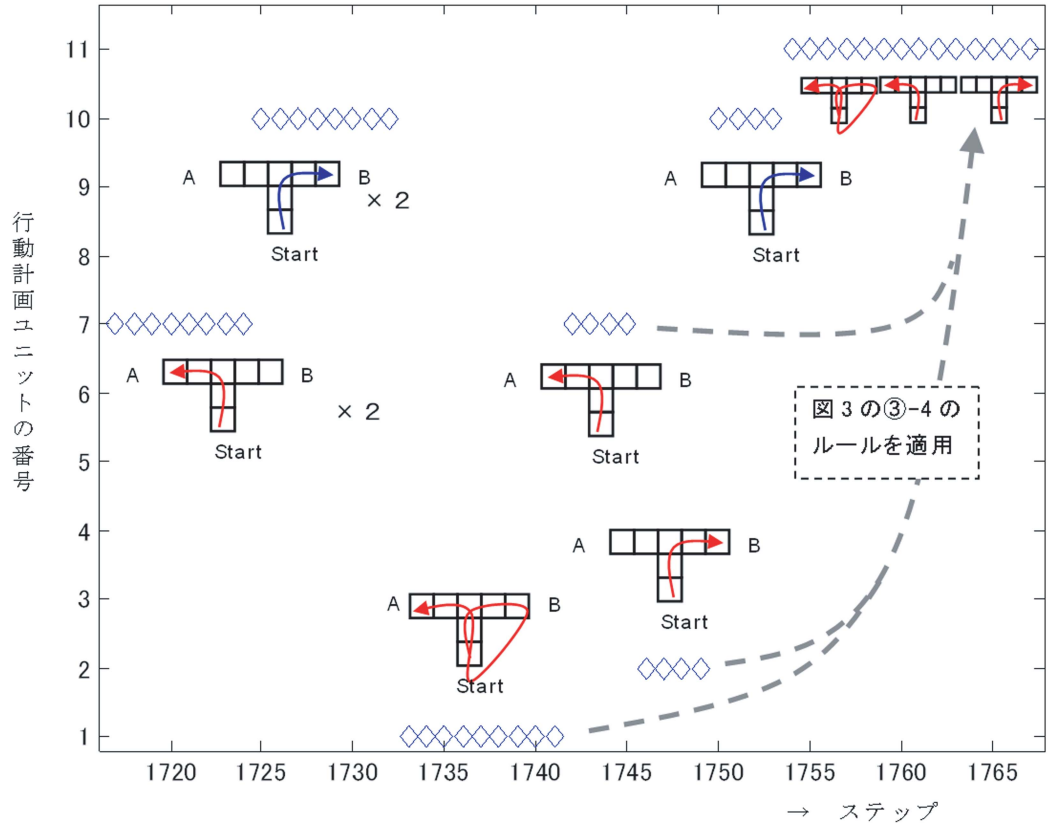

図 10 予備段階から本課題への移り変わり（図 9 B の 1750 ステップ付近）の拡大図.
青の菱形は各ステップにおいて選択された行動計画ユニット．T 字型迷路におけるエージェントの実際の行動も示した．赤の矢印は成功したときの行動経路で，青の矢印は失敗したときの行動経路を示す．11 番目の行動計画ユニットは，グレーの点線で示された 3 つの行動計画ユニットの連結によって作成された．1-7-2 と 11 番目の行動計画ユニットとでは全体として同じ行動系列でも，課題の変化に柔軟に対応できるよう行動計画ユニットは別に作成される．

Fig. 10 Enlarged figure around the 1750th step of Fig. 9 B.
Blue lozenges indicate Action plan units selected at each step. The agent's behaviour are also shown. Red arrows show successful behaviour, and blue arrows show failure behaviour. The 11th Action plan unit was made by the integration of three Action plan units as shown with gray arrows. Although the 11th Action plan corresponds with the connection of 1, 7 and 2 Action plan units, Action plan unit is made independently in order to adapt for task's change.

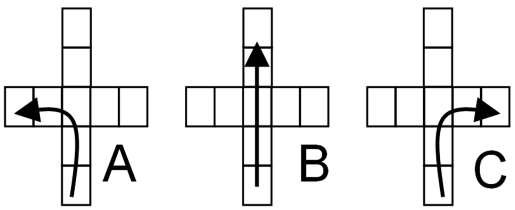

図 11 十字型迷路で用いられる 3 つの試行 A，B および C.
本課題：“AAABBCCCCBABCABCCABC” の繰返し（1 周期 20 試行）

Fig. 11 Trials A, B and C of the cross maze task.
Main task: “AAABBCCCCBABCABCCABC” is repeated. One period consists of 20 trials.

み」よりもステップ数が少なく，ゴールの位置変化のルールが複雑な場合も過去の経験が役立っていることが分かる．

なお，**図 14** に十字型迷路の本課題のみにエージェントが取り組んだときの典型的な結果を示した．行動計画時に P 層で選択された行動計画ユニットの推移を示しており，横軸は試行数で縦軸は何番の行動計画ユニットが選択されたかを表している．ただし縦軸の 0 番目は探索行動を意味する．

図 14 によると，エージェントは行動計画ユニットを次々と増やして経験を積みつつも，行動計画時に選

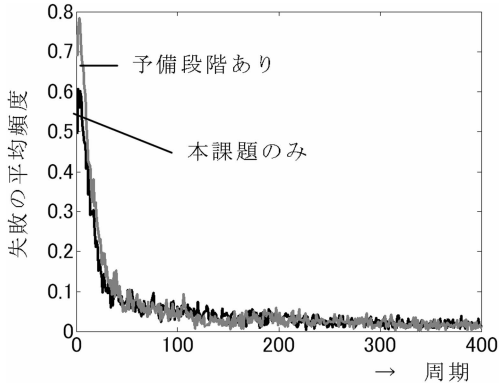

図 12 十字型迷路の学習曲線：複雑なルールも自律的に獲得できている
Fig. 12 Learning curve of the cross maze task: Agents could learn the complicated rule autonomously.

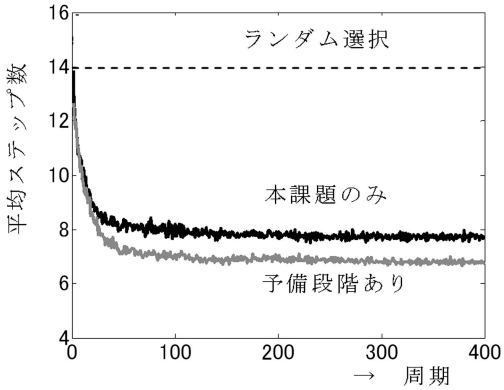

図 13 十字型迷路の各周期の 1 試行あたりの平均ステップ数 ルールが複雑な場合も過去の経験が役立っている
Fig. 13 Average steps per one trial in the cross maze task: Experience of pre-learning was available for solving the complicated rule.

択される行動計画ユニットの範囲は比較的最近のものにしぼりながら学習を進めていることが分かる．そして最終的には，エージェントは複数（この例では 3 つ）の行動計画ユニットの組合せとして解を出している．

4. 考　　察

本論文ではまず，ゴールの位置がある規則で変化する場合のように，POMDPs 下の環境でもエージェントが自律的に時間を区切らないと解けない課題があるとした．そしてそのような課題の場合，Profit sharing やメモリベース法，確率的政策などの従来手法では解決に困難をともなうと位置づけた．さらに Profit sharing で学習に利用されるエピソードの枠組みを拡張し，エピソードの貯蔵・編集によって行動計画を立てる学習アルゴリズムを提案した．

提案アルゴリズムを迷路課題に適用した結果，Profit sharing と比べて環境変化に頑健で（図 7），ゴールの位置の変化則が長い場合でも次のゴールを誤りなく予測できたこと（図 12）が確認された．また，予備段階の学習が本課題でより良い答えを見つけるのに役立っていることを示し（図 7，図 13），学習の干渉よりむしろ積み重ねの効果があることが示唆された．

そのため，課題の変化が積み重ねとしてかえって学習にプラスの効果をもたらす場合を学習の積み重ね効果として次の 4.1 節で考察する．また，本研究では状態遷移が非決定的な環境は扱わなかったし，解もランダム選択より優れていたとはいえ最適解を得られたわけではなかった．これらの点について，それぞれ 4.2 節と 4.3 節で考察を行う．

4.1 学習の積み重ね効果

T 字型および十字型迷路課題において，予備段階で試行 A を経験した後に正反対の試行 B を経験しても，エージェントは学習の干渉をあまり示さなかった．課題が変化した場合に学習の干渉をなるべく抑え，逆に過去の経験が新しい課題にも利用できうるようにするには，1）個々のエピソードを別個に保存することと，2）それらのエピソードの中でも有効なものを確実に選択することが必要である．

本アルゴリズムでは，新しい経験と見なされたエピソードを行動計画ユニットとして個別に保存し，さらに行動計画ユニット間に成功・失敗に応じた結合関係を形成していた（図 3 の ③）．また，各行動計画ユニットに付与された内的評価値をもとにして，適切な行動計画ユニットの選択に役立てていた（図 3 の ①）．これらが前述の 2 点を実現させ，干渉を抑えていたといえる．

学習の積み重ね効果は，より複雑な課題への取り組みや学習の促進に重要である．より複雑な課題としてたとえば，T 字型迷路の本課題として AAB を 7 周期後，BBA を 7 周期の繰返し（1 周期 42 試行）のように 2 重の周期性がある場合が考えられる．この場合には AAB および BBA の 3 試行周期と AAB × 7 回，BBA × 7 回が交互に続く 42 試行周期の 2 重の周期性がある．

紙数の都合で図示しないが，予備段階として AAB と BBA を各 200 周期ずつ経験させた後に本課題を行わせた場合，同じ周期 42 試行でも 2 重の周期性を示さない課題を予備段階なしで行わせた場合と比べ，エージェントはずっと早く学習を収束させた．予備段階を経たエージェントは，本課題において 3 試行ごとに行動を区切ることで効率良く解を見い出していた．

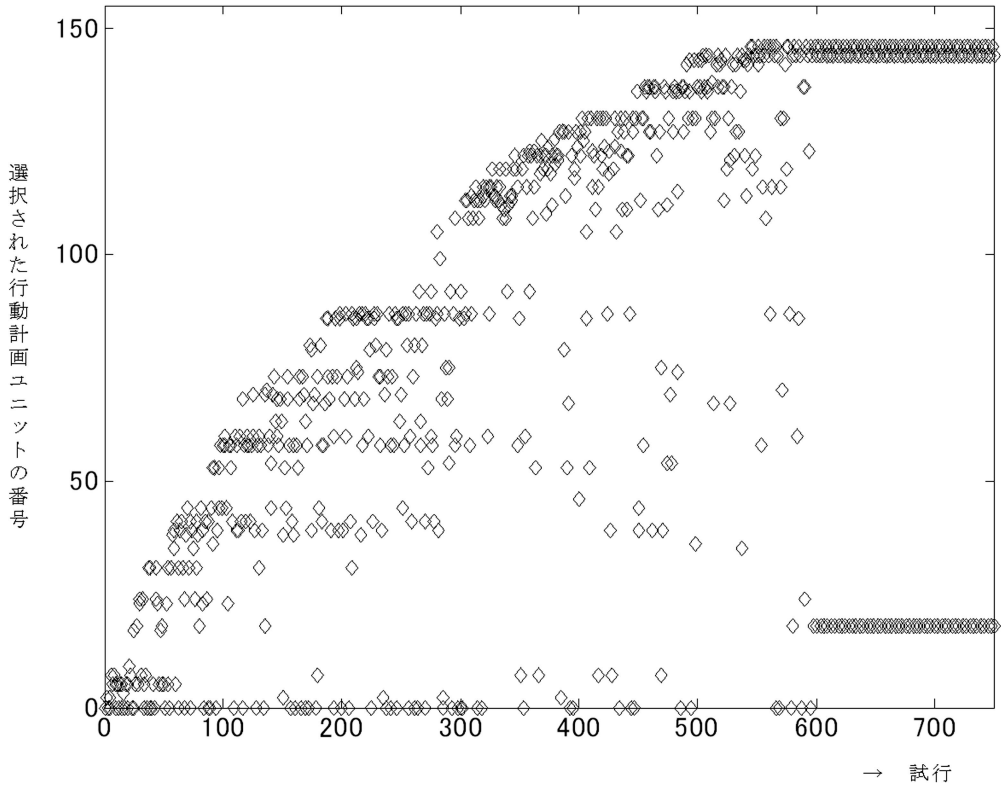

図 14 十字型迷路の本課題のみを解いたときの P 層のダイナミクス例.
各試行の行動計画で選択された行動計画ユニットを菱形で示す．エージェントは行動計画ユニットの数を増やして経験を積みつつも，選択される行動計画ユニットをしぼり，最終的に 3 つの行動計画ユニットの組合せとして解を出している

Fig. 14 Example of the P layer's dynamics when the agent learned the cross maze task.
Lozenges indicate Action plan units selected at each trial. The agent increased the number of units, and narrowed candidates. Finally, the agent solved the task as a combination of three Action plan units.

今後，他の実験条件も加えて検討し，学習の積み重ね効果を実現するより詳細な仕組みを明らかにしたい．

4.2 非決定的な状態遷移が存在する環境

本研究で取り扱った問題は，報酬の関数（ゴールの位置）に不完全知覚が存在する部分観測マルコフ決定過程である．しかしこれまでの課題における状態遷移は決定的であり，非決定的な環境ではなかった．提案アルゴリズムは過去の経験をエピソードとして貯蔵し，それらの組合せや統合による編集を行うことで隠れたルールを見い出す．しかし，たとえば状態遷移がランダムに変わるような，非決定的な状態遷移が存在する課題では，エピソードの作成を無限に繰り返してしまうであろう．

そのため，たとえば「左に曲がったつもりが右に来てしまった場合」など確率的遷移によって予想外の地点に来てしまった状況を 1 つのエピソードとして学習できれば，より適応的な行動を示せるのではないか．非決定的な状態遷移も含めた広範な課題を取り扱えるようにするための学習原理や機能を付加することが今後の課題といえる．

4.3 解の最適性

本アルゴリズムの実現にあたっては，4.1 節で述べたように，過去のエピソードが新規の課題に利用されるよう個々のエピソードを独立に保存し，後で想起・編集できるようにする必要があった．そのためゴールまでの最短経路は保証できなかった．提案アルゴリズムは環境の変化による学習の干渉をあまり示さなかったので，予備段階の課題のパターンを増やすことで改善が期待できる．しかし T 字型迷路の「予備段階あり」で AAA... と BBB... のほかに ABABAB... も予備段階として加えたところ，若干の改善がみられたが最短ステップには至らなかった．おそらくは，ショート

カットの実現が促される何らかの仕組みを付与する必要がある．

Aota ら[9] はエージェントの選択した行動を状態として考慮することで，新規の環境でも 1 度ゴールを経験しただけでショートカットが実現できる学習モデルを提案した．Aota らのモデル，または従来の強化学習モデルと本アルゴリズムを組み合わせることで，エージェントが安定的に最短経路を選択できるかもしれない．また，「『遠回りな経路』から『最短経路』を選んだ経験」を 1 つのエピソードとして，エージェントが学習できるようアルゴリズムを改良できないかも検討したい．

5. ま と め

環境変化の規則も学習できるようにするために，エピソード記憶編集に基づいた新しい学習アルゴリズムを提案した．このアルゴリズムを実装したエージェントを Alternation maze task に適用した結果，エージェントは部分観測マルコフ決定過程下でもとくに時間軸上の不確定さが存在する課題に対して適切なエピソードを選択できた．また，過去に学習した状況も新奇な別の状況に対してより良い解の探索に役立っていることが分かった．

今後の課題として，1）学習の積み重ね効果の検証，2）非決定的な状態遷移が存在する課題への拡張，3）最短経路の保証などがあげられた．

謝辞 採録にいたるまでに辛抱強くお世話くださった編集委員の諸先生方，ならびに本論文の査読者に深く感謝します．査読者の指摘により，本論文を大きく改善することができました．また，有益なコメントを下さった MPS 研究会の諸先生方にも感謝の意を表します．

参 考 文 献

1) 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，2003 年度人工知能学会全国大会（第 17 回）講演論文集，Vol.3F4-03 (2003).
2) Wood, E.R., Dudchenko, P.A., Robitsek, R.J. and Eichenbaum, H.: Hippocampal Neurons Encode Information about Different Types of Memory Episodes Occurring in the Same Location, *Neuron*, Vol.27, pp.623–633 (2000).
3) 山口陽子：海馬の動的神経機構を基礎とする状況依存的知能の設計原理，CREST「脳を創る」平成 11 年度採択課題，独立行政法人科学技術振興機構 (1999).
4) 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，人工知能学会論文誌，Vol.18, pp.286–296 (2003).
5) McCallum, R.A.: Instance-Based Utile Distinction for Reinforement Learning with Hidden State, *Proc. 12th International Conf. on Machine Learning*, pp.387–395 (1995).
6) 木村 元, Kaelbling, L.P.：部分観測マルコフ決定過程下での強化学習，人工知能学会誌，Vol.12, pp.822–830 (1997).
7) 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, pp.104–111 (1994).
8) 鈴木淳司，松井藤五郎，世木博久：罰を考慮した profit sharing 強化学習法，2003 年度人工知能学会全国大会（第 17 回）講演論文集，3F4-02 (2003).
9) Aota, Y. and Miyake, Y.: Neurogenesis performs the formation of the cognitive space in Rat's navigation, *Neurocomputing*, Vol.62, pp.161–178 (2004).

付 録

学習アルゴリズムの数式表現

学習アルゴリズムを数式で表したものを以下に示す．ここで e_i は i 番目の行動計画ユニット，n_E はエージェントが貯蔵した行動計画ユニットの総数，p_E は行動計画ユニットの番号で開始時における 1 つ前の試行の行動計画ユニットを示す．$w_{j,i}^{EE}$，w_i^{ER}，b_i，I_i はそれぞれ i 番目の行動計画ユニットの，j 番目の行動計画ユニットからの結合強度，内的評価値，報酬の数，経験値を示す．p_S はエージェントの現在位置（知覚入力）の番号であり，$w_{j,i}^{SE}$ は j 番目の位置から i 番目の行動計画ユニットの結合強度を表す．γ は 0—1 の間の乱数，ε は探索行動の度合いを決めるための閾値で 0.3 とした．

0. 初期状態

$$n_E = 1$$
$$p_E = 0$$
$$w_{p_E,1}^{EE} = 0$$
$$w_1^{ER} = 0$$
$$b_1 = 0$$
$$I_1 = 0.1$$

1. 開始時の行動決定

$$r_i = w_{p_E,i}^{EE}(w_i^{ER} + 0.1\gamma)$$
$$w_i^{ER\prime} = w_i^{ER}, \quad i = 1, \ldots, n_E$$
$$p_E' = \arg \max_{i=1,\ldots,n_E} r_i$$

$$\begin{cases} s=1, & r_{p'_E}>0 \\ & \wedge\, b_{p'_E}>0 \\ & \wedge\, r_{p'_E} I_{p'_E} > \gamma \\ s=0, & \\ w^{ER\prime}_{p'_E}=-1, & \text{otherwise} \end{cases}$$

$s=1$ のとき，$e_{p'_E}$ が選択され 2 へ．
$s=0$ のとき，下記に従う．

$$\begin{cases} r_i = -w^{SE}_{p_S,i}, & w^{EE}_{p_E,i}<0 \\ r_i = w^{SE}_{p_S,i} w^{ER\prime}_i \gamma, & \text{otherwise} \end{cases} \qquad i=1,\ldots,n_E$$

$$p'_E = \arg\max_{i=1,\ldots,n_E} r_i$$

$$p'_E = 0, \quad r_{p'_E}\gamma < \varepsilon \vee b_{p'_E} = 0$$

2. 選択行動計画ユニットの実行，新しい行動計画ユニットの作成と編集，および行動計画ユニット間の結合

A）$p'_E=0$ のとき **B）**へ．$p'_E>0$ のとき，下記に従う．

(1) $e_{p'_E}$ 中の成功エピソードと同じ行動を実行する．

(2) $e_{p'_E}$ の実行中，$e_{p'_E}$ とは異なる時点で（予想外の）報酬を得たときは (5) へ．

(3) $e_{p'_E}$ の実行中，$e_{p'_E}$ のときは得られた報酬を（予想がはずれ）得られなかったとき（これを失敗と呼ぶことにする）は (6) へ．

(4) $e_{p'_E}$ における最後の報酬まで（予想どおり）実行できたときは

$$I_{p'_E} \leftarrow I_{p'_E} + 0.3$$
$$I_{p'_E} = 1, \quad I_{p'_E} > 1$$
$$w^{EE}_{p_E,p'_E} = 1$$
$$p_E = p'_E$$

とし，試行終了（次の試行に移る場合，(1) へ）．

(5) $w^{EE}_{p_E,p'_E} = -1$

- $e_{p'_E}$ として実行し始めてから最後に（予想外の）報酬を得た時点までの行動が，別の行動計画ユニット $e_{p''_E}$ の行動と同じとき，

$$p'_E = p''_E$$
$$w^{EE}_{p_E,p'_E} = 1$$
$$p_E = p'_E$$

とし，試行終了（次の試行に移る場合，(1) へ）．

- $e_{p'_E}$ として実行し始めてから最後に（予想外の）報酬を得た時点までの行動と同じ行動計画ユニットがないとき，

$$n_E \leftarrow n_E + 1$$
$$p'_E = n_E$$

新しい行動計画ユニット $e_{p'_E}$ としてこの行動を貯蔵．

$$w^{EE}_{p_E,p'_E} = 1$$
$$I_{p'_E} = 0.1$$
$$w^{ER}_{p'_E} = 4\left(b^2_{p'_E} \Big/ t_{p'_E}\right)$$

ここで $b_{p'_E}$ は $e_{p'_E}$ で得た報酬の回数，$t_{p'_E}$ は $e_{p'_E}$ における総ステップ数である．

$$p_E = p'_E$$

とし，試行終了（次の試行に移る場合，(1) へ）．

(6) $w^{EE}_{p_E,p'_E} = -1$
$I_{p'_E} = 0.1$

a) $e_{p'_E}$ として実行し始めた今回の行動と同じ行動を示す行動計画ユニットがないときは b) へ．別の行動計画ユニット $e_{p''_E}$ の行動と同じとき，

$$p'_E = p''_E$$
$$w^{EE}_{p_E,p'_E} = 0$$

1 つ前の失敗から今回の失敗の間（今回が初めての失敗の場合は，1 回目の試行から今回の失敗の間）に 2 回以上 30 回以下まで報酬を得た連続成功エピソードがあるときは c) へ．ない場合は，

$$p_E = p'_E$$

とし，(1) へ．

b) $n_E \leftarrow n_E + 1$
$$p'_E = n_E$$

新しい行動計画ユニット $e_{p'_E}$ としてこの行動を貯蔵．

$$w^{EE}_{p_E,p'_E} = 0$$
$$w^{ER}_{p'_E} = 4\left(b^2_{p'_E} \Big/ t_{p'_E}\right) + (2/t_{\text{fail}})$$

ここで $b_{p'_E}$ は $e_{p'_E}$ で得た報酬の回数，$t_{p'_E}$ は $e_{p'_E}$ において最後に報酬を得たときまでのステップ数，t_{fail} は $e_{p'_E}$ で最後に報酬を得てから失敗したときまで

のステップ数である．
1 つ前の失敗から今回の失敗の間（今回が初めての失敗の場合は，1 回目の試行から今回の失敗の間）に 2 回以上 30 回以下まで報酬を得た連続成功エピソードがあるときは c) へ．ない場合は，

$$p_E = p'_E$$

とし，(1) へ．

c) 2 回以上報酬を得た連続成功エピソードを 1 つの行動計画ユニットと見なし，それが別の行動計画ユニット $e_{p''_E}$ として経験したのと同じときは d) へ，新しい経験のときは e) へ．

d) p_E^f を 1 つ前の失敗の行動計画ユニットの番号とする．

$$p'_E = p''_E$$

$$w_{AB}^{EE} = 1, \quad (A \equiv p_E^f,\ B \equiv p'_E)$$

$p_E = p'_E$ とし，(1) へ．

e) p_E^f を 1 つ前の失敗の行動計画ユニットの番号（ない場合は 0）とする．

$$n_E \leftarrow n_E + 1$$

$$p'_E = n_E$$

$$\begin{cases} w_{AB}^{EE} = 1, & A > 0 \\ w_{AB}^{EE} = 0, & A = 0 \end{cases}$$

$$(A \equiv p_E^f,\ B \equiv p'_E)$$

$$I_{p'_E} = 0.1$$

$$b_{p'_E} \equiv \left\{ e_{p'_E} \text{ で得た報酬の回数} \right\}$$

$$t_{p'_E} \equiv \left\{ e_{p'_E} \text{ における総ステップ数} \right\}$$

$$w_{p'_E}^{ER} = 4\left(b_{p'_E}^2 \Big/ t_{p'_E} \right)$$

$p_E = p'_E$ とし，(1) へ．

B) ランダムな行動（探索）を実行する．
報酬を得たとき，探索開始から報酬を得るまでの行動がそれまで貯蔵したどのエピソードの行動とも異なるものである場合，（イ）へ．
$e_{p''_E}$ と同じ行動の場合は（ロ）へ．

（イ） $n_E \leftarrow n_E + 1$

$$p'_E = n_E$$

$$\begin{cases} w_{p_E,p'_E}^{EE} = 1, & p_E > 0 \\ w_{p_E,p'_E}^{EE} = 0, & p_E = 0 \end{cases}$$

$$I_{p'_E} = 0.1$$

$$b_{p'_E} \equiv \left\{ e_{p'_E} \text{ で得た報酬の回数} \right\}$$

$$= 1$$

$$t_{p'_E} \equiv \left\{ e_{p'_E} \text{ における総ステップ数} \right\}$$

$$w_{p'_E}^{ER} = 4\left(b_{p'_E}^2 \Big/ t_{p'_E} \right)$$

$p_E = p'_E$ とし，試行終了（次の試行に移る場合，(1) へ）．

（ロ） $p'_E = p''_E$

$$\begin{cases} w_{p_E,p'_E}^{EE} = 1, & p_E > 0 \\ w_{p_E,p'_E}^{EE} = 0, & p_E = 0 \end{cases}$$

$p_E = p'_E$ とし，試行終了（次の試行に移る場合，(1) へ）．

(平成 17 年 8 月 20 日受付)
(平成 17 年 12 月 25 日再受付)
(平成 18 年 1 月 20 日採録)

青田 佳士（正会員）

昭和 45 年生．平成 12 年東京工業大学大学院情報理工学研究科数理・計算科学専攻博士後期課程単位取得退学．同年科学技術振興機構戦略的基礎研究推進事業『脳を創る』山口チーム研究員．理化学研究所脳科学総合研究センター創発知能ダイナミクス研究チーム非常勤研究員．平成 16 年横浜国立大学大学院国際社会科学研究科助手．エピソード記憶による学習理論の研究に従事．情報処理学会数理モデル化と問題解決研究会，日本神経回路学会各会員．

山口 陽子

昭和 30 年生．昭和 56 年東京大学大学院薬学系研究科製薬化学専攻博士課程中退．昭和 57 年東京大学薬学部教務職員．昭和 59 年東京大学薬学部助手．平成 5 年東京大学薬学部講師．同年東京電機大学理工学部情報科学科助教授．平成 8 年東京電機大学理工学部情報科学科教授．平成 12 年より理化学研究所脳科学総合研究センター脳型知能システム研究グループ創発知能ダイナミクス研究チームチームリーダー．また，兼任として平成 11 年科学技術振興機構戦略的基礎研究推進事業『脳を創る』「海馬の動的神経機構を基礎とする状況依存的知能の設計原理」研究代表者．平成 12 年東京電機大学大学院理工学研究科客員教授（生物情報論）．同年東京電機大学大学院理工学研究科講師（生物情報論）．平成 15 年東京大学理学部講師（生物情報プログラム：数理神経生物学）．同年東京大学工学部講師（脳科学入門）．計算論的神経科学の研究に従事．薬学博士．平成 6 年度日本神経回路学会論文賞受賞．日本生物物理学会，日本物理学会，日本神経回路学会，日本神経科学会，電子情報通信学会，計測自動制御学会，ニューロエソロジー学会，数理生物学会，Society for Neuroscience 各会員．