論 文

Technical Papers

罰を回避する合理的政策の学習

Reinforcement Learning for Penalty Avoiding Rational Policy Making

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, National Institution for Academic Degrees
teru@niad.ac.jp

坪井 創吾
Sougo Tsuboi
(株) 東芝 研究開発センター ヒューマンインターフェースラボラトリー
Human Interface Laboratory, Corporate Research & Development Center, TOSHIBA
sougo.tsuboi@toshiba.co.jp

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, reward and penalty, penalty avoiding, rational policy making

Summary

Reinforcement learning is a kind of machine learning. It aims to adapt an agent to a given environment with a clue to rewards. In general, the purpose of reinforcement learning system is to acquire an optimum policy that can maximize expected reward per an action. However, it is not always important for any environment. Especially, if we apply reinforcement learning system to engineering, environments, we expect the agent to avoid all penalties.

In Markov Decision Processes, a pair of a sensory input and an action is called rule. We call a rule *penalty* if and only if it has a penalty or it can transit to a *penalty state* where it does not contribute to get any reward. After suppressing all penalty rules, we aim to make a rational policy whose expected reward per an action is larger than zero. In this paper, we propose a *suppressing penalty algorithm* that can suppress any penalty and get a reward constantly. By applying the algorithm to the tick-tack-toe, its effectiveness is shown.

1. はじめに

強化学習とは, 報酬という特別な入力を手がかりに環境に適応した行動決定戦略を追求する機械学習システムである. そこでは, 「何をして欲しく, 何をして欲しくないか (what)」という学習目標を報酬に反映させるだけで, 「その実現方法 (how to)」を学習システムに獲得させることができる.

このため強化学習では, 報酬をどのように設計するかが非常に重要な問題となる. 近年, 多くの強化学習研究では, 目標達成時に正の報酬, 制約違反時に罰として負の報酬を付与し, 単位行動当たりの期待獲得報酬の最大化, すなわち**最適政策**の獲得を追求する場合が多い [Sutton 98]. しかし与えられた報酬および罰の値によっては, 学習された最適政策が設計者の期待に反する可能性がある.

例えば, 将棋などの対戦型ゲームにおいて, 勝ちに対し正の報酬, 負けに対し負の報酬を与えた場合, 必勝戦略が存在したとしても, それらの値によっては負ける可能性のある戦略を学習する場合がある. 具体例として,100 手で完了する必勝戦略と, 10 手で完了する勝率 90%の戦略を比べてみる. 勝ちに対し 100, 負けに対し -100 の報酬を与えた場合, 単位行動当たりの期待獲得報酬量は前者で $1 = (\frac{100}{100})$, 後者で $8 = (\frac{100*0.9+(-100)*0.1}{10})$ となり, 負ける可能性のある後者の方がよい値を示す. 未知環境を対象とする以上, 正および負の報酬の設計に起因するこのような問題は, つねに生じる可能性がある.

さらに, 最適政策は, 実際上は必ずしも有効であるとは限らない. 特に大規模問題などでは, その獲得の容易さから, 最適政策よりも, 単位行動当たりの期待獲得報酬量が正である**合理的政策**の方が重視される場合も多いと考える.

本論文では, 目標達成時に与えられる**報酬**と制約違反時に与えられる**罰**を区別して取り扱うことで上記の問題を解決する. 罰を得ることのない政策を**罰回避政策**と呼び, 合理的政策とは独立に評価する. そのため, 合理的政策では罰を無視し, 罰回避政策では報酬を無視する点に注意されたい.

まず, 我々は, 強化学習を行う以上, 報酬を得なければ意味がないと考える. すなわち合理的政策の獲得は必須である. この前提の下で, 罰回避政策の中に合理的政策が

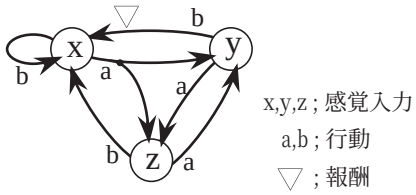

図 1 報酬の存在する環境の例:E1.

存在するときは, その中で, できるだけ安定して多くの報酬が得られる政策を学習することが望ましい. 一方, 罰回避政策の中に合理的政策が存在しないときは, 罰を得る可能性が低い合理的政策を学習することが望ましい. 本論文では, 報酬および罰にどのような値が設定されていたとしても, このような望ましい学習をつねに実現する手法を提案する.

以下, 第 2 章では, 本研究における問題設定を述べる. 第 3 章では, 任意の報酬および罰の値の下で, 極力罰を排除し, つねに継続して報酬を得るための手法である**罰回避政策形成アルゴリズム**を提案する. 第 4 章では, 従来手法との比較を行なった後, 代表的なゲーム問題として 3 目並べの一種である tick-tack-toe を取り上げ, 提案手法を評価する. 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 対 象 問 題

本論文では, **離散マルコフ決定過程** (MDPs) 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**または**罰**が与えられる. 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. 各状態に対し, 選択すべきルールを与える関数を**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**, 罰を得ることのない政策を**罰回避政策**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. 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, zb, xa, yb, (報酬)] と行動したとすると, このなかには (xb, xa, ya, za, yb), (xa, zb, xa, 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.

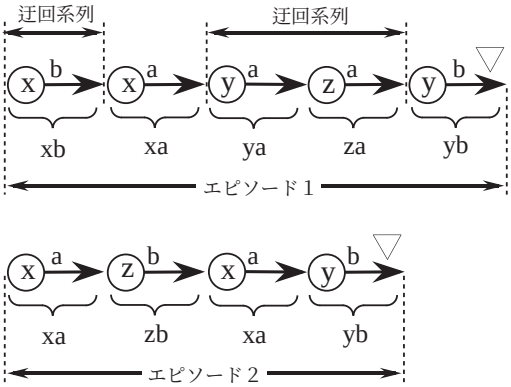

図 2 環境 E1 におけるエピソードと迂回系列の例.

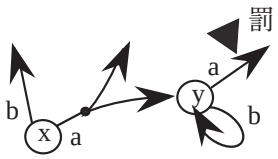

図 3 罰ルール (xa,ya) および罰状態 (y) の例.

ある合理的政策の構成要素となっているルールを **合理的ルール**, それ以外のルールを **非合理的ルール** と呼ぶ. 図 1 の環境では, xa,yb,za,zb の 4 つが合理的ルールであり, xb と ya が非合理的ルールである [宮崎 99a]. あるエピソードにおいて迂回系列上に存在しないルールは合理的である. 図 2 では xa,yb,zb がこれに該当する. 迂回系列上に存在していても, 同じエピソードにおいて同一の感覚入力に対して異なるルールが存在しなければ, そのルールは合理的である. 図 2 では za がこれに該当する.

さらに,「直接罰を得た経験のあるルール」または「選択可能なルールが, 罰ルールまたは非合理的ルールのみである遷移先を持つルール」を**罰ルール**と呼ぶ. 選択可能なルールが罰ルールまたは非合理的ルールのみである状態を**罰状態**と呼ぶ. 例えば, 図 3 ではルール xa と ya は罰ルールであり, 状態 y は罰状態である.

本論文では, 任意の報酬および罰の値の下で, 罰状態への遷移をできるだけ回避する合理的政策の形成を目標とする.

2·2 問 題 の 所 在

Q-learning(QL)[Watkins 92] や Policy Iteration Algorithm(PIA)[ワグナー 78] では, 報酬に正のある値, 罰に負のある値を付与することで, 割引期待獲得報酬を最大化する政策である**最適政策**を求めることが可能である. しかしこの場合, 先に述べた学習目標を実現する報酬および罰の値の設計は, 一般には, 容易ではない.

例えば, 図 4 に示す環境 E2 を考える. 本論文の立場から, 罰 (P) を完全に回避する合理的政策の獲得が最優先とされるが, そのような政策が存在しない場合には, 極力

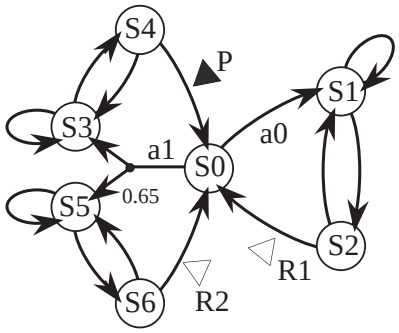

図 4 報酬と罰が存在する環境の例:E2. 状態 S0 で行動 a1 を実行した結果, 状態 S3 および S5 へ遷移する確率は, それぞれ 0.35 と 0.65 である.

罰を得ない合理的政策の獲得を目指すこととなる. この目標を達成するための報酬 (R1,R2) および罰 (P) の設計方針としては, R1 および R2 に正のある値を付与し, P に負の大きな値 (例えば $P = -10000$ 等) を付与するというものが考えられる.

しかし強化学習は未知環境を対象とする学習手法であり, 一般には, このような形で環境の完全なる構造が与えられるとは限らず, 先の設計方針は容易に破綻をきたす. 例えば, 同じ環境の構造であっても,R1 が存在しない場合には, 罰を大きくし過ぎると, S0 で a0 を選択する報酬の得られない政策が形成される. また,R1 が存在したとしても, 例えば,S0 から S3 への遷移確率が小さい場合などでは, 罰を完全に回避する合理的政策が存在するにも関わらず, S0 で a1 を選択する罰を得る可能性のある政策が形成される.

すなわち, 未知環境を対象とする以上, 一般的な報酬および罰の値の設計方針を与えることは非常に困難である.

2·3 接 近 法

本論文では, 合理的政策の獲得を前提とした上で, 報酬と罰の取り扱いについて次のように定める.

(1) 罰回避が可能な場合, 罰回避政策の中で, 各状態で最悪の状態遷移をした場合に最も少ない行動数で報酬を得ることが期待できる合理的政策を学習する.

(2) 罰回避が不可能な場合, 罰状態への遷移をできるだけ低くする合理的政策を学習する.

任意の報酬および罰の値の下で, 上記を実現する手法として**罰回避政策形成アルゴリズム**を次章で提案する.

3. 罰回避政策形成アルゴリズムの提案

3·1 基 本 方 針

本論文では, 罰ルールを手掛かりに罰を極力回避する合理的政策の形成を目指す. 罰を回避するためには, まずルール集合の中から罰ルールを排除することが必要である. 3.2 節では, そのための手法として**罰ルール判定手続き**を提案する. 罰ルールが排除された後には, 合理的政策をどのようにして形成するかが問題となる. 3.3 節では, 罰回避政策の中に合理的政策が存在する場合と存在しない場合に分けて議論する. 以上を統合した手法として**罰回避政策形成アルゴリズム**を 3.4 節で提案し, その特徴を 3.5 節でとりまとめる.

procedure 罰ルール判定手続き
begin
これまでに経験したエピソードの中で,
直接罰を得たことにのあるルールにマークする.
do
以下の条件が成立する状態にマークする.
その状態で選択可能なルールが非合理的ルールまたは
マークが付与されたルールのみである.
以下の条件が成立するルールにマークする.
そのルールで遷移可能な状態の中の少なくとも
ひとつがマークされている.
while 新たに少なくともひとつの状態がマークされる.
end.

図 5 罰ルール判定手続き.

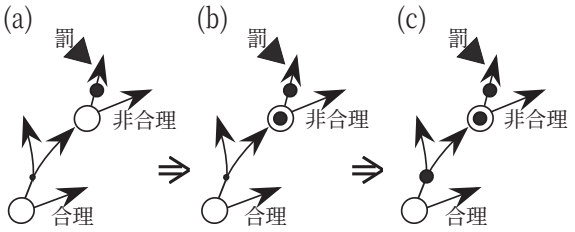

図 6 罰ルールの判定.

3·2 罰ルール判定手続き

エピソードごとに図 5 に示す罰ルール判定手続きを適用する. これはルール集合の中から, その時点でのすべての罰ルールを発見するためのアルゴリズムであり, 罰ルールを判定するために, マークを利用する. これまでに経験したエピソードの中で, 直接罰を得たことのあるルールにマークする (図 6a). その後, その状態で選択可能なルールが非合理的ルールまたはマークが付与されたルールのみである状態にマークする (図 6b). さらに遷移可能な状態の中の少なくともひとつがマークされているルールにマークする (図 6c). この結果, マークが付与されたルールは罰ルールとなる. 以上の手続きを状態にマークが付与されなくなるまで繰り返す. 上記手法により, その時点での罰ルールをすべて発見することができる.

マークを正しく伝播させるためには, 各ルールの遷移先状態をすべて記憶しておく必要がある. そのため, 状態の種類を n, 行動の種類を m とした場合, 空間計算量は $\mathrm{O}(mn^2)$ となる. ただし, ゲームのようにルールが既知な場合は, 各ルールの遷移先状態が計算可能なので, 空間計算量は $\mathrm{O}(mn)$ となる.

procedure 合理的政策改善手続き
begin
直接報酬を得たことのあるルールを政策に取り込み,
そのルールを選択可能とする状態にマークする.
do
以下の条件が成立するルールを政策に取り込む.
その状態での政策が未定であり，かつ，その
ルールで遷移可能な状態全てがマークされている.
以下の条件が成立する状態にマークする.
その状態で選択すべき政策が決定されている.
if 政策がひとつも決定されない then
任意のマーク伝播停止ルールを政策に取り込み,
そのルールを選択可能とする状態にマークする.
while 新たに少なくともひとつの状態がマークされる.
end.

図 7 合理的政策改善手続き.

3·3 罰を回避する合理的政策の形成

罰回避政策の中に合理的政策が存在する場合には, 罰ルール判定手続きにより罰ルールを判定した後に, ルール集合内からそれらの罰ルールをすべて取り除き, 合理的政策の形成を行うことができる. そのための手法について§1 で述べる. 一方, 罰回避政策の中に合理的政策が存在しない場合には, §1 で述べた手法の中に, 罰を確率的に回避する機構を組み込む必要がある. そのための手法について§2 で述べる.

§1 合理的政策改善手続き〜罰回避政策の中に合理的政策が存在する場合〜

政策形成アルゴリズムとして PIA がよく知られている. 罰ルールが予め排除されていれば,PIA を利用することにより, 罰ルールを除いた空間における最適政策を得ることができる. PIA は多項式時間で実行可能な手法である [Papadimitriou 87] が, 一般に, 膨大な状態を持つ問題に適用することは困難である. ここでは,PIA を利用せずに政策を改善することが可能な手法として**合理的政策改善手続き**を提案する.

図 7 に示す提案手法では, 報酬を直接得たことのあるルールからマークを伝播させることで徐々に政策を形成していく. まず, 直接報酬を得た経験のあるルールを政策に取り込み, そのルールを選択可能とする状態にマークする (図 8a). その後, その状態での政策が未定であり, かつ, そのルールで遷移可能な状態すべてがマークされているルールを政策に取り込み, その状態にマークする (図 8b). 以上を状態間でマークの伝播が停止するまで繰り返す.

報酬を得るために必要なルールであっても, 状態遷移の非決定性のため, ある特定の遷移先がマークされない場合も考えられる. 例えば, 図 9 では, 状態 y にマークが伝播されたとしても, 状態 x がマークされない限り状態 z はマークされないので, ルール xa が政策に取り込まれることがなく, 状態 y でマークの伝播が停止してしまう. 以下ではそのようなルールを**マーク伝播停止ルール**と呼ぶ. マーク伝播停止ルールは報酬を得るために必要なルールである. そこで, 完全にマークの伝播が停止した後に, マーク伝播停止ルールの中から任意のひとつを政策に取り込む. そしてそのルールを選択可能とする状態にマークし, 再び, マークの伝播を試みる. 以上をマークの伝播が停止するまで繰り返す.

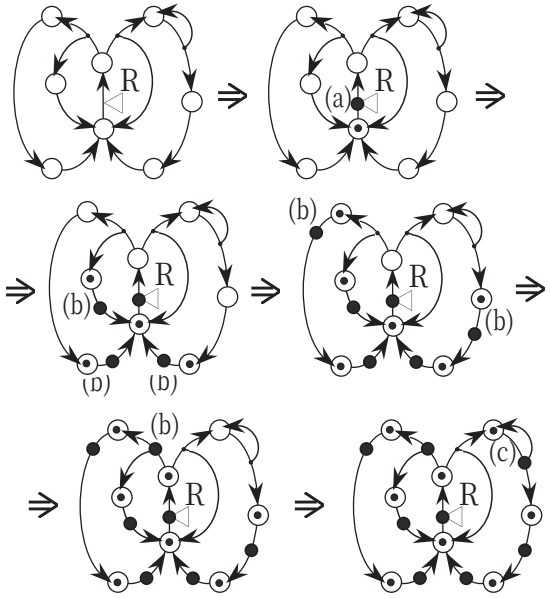

図 8 合理的政策改善手続きの動作例.

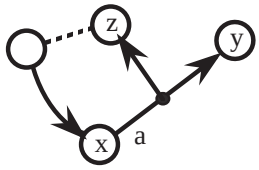

図 9 マーク伝播停止ルールの例.

マーク伝播停止ルールは, 最悪の場合, 報酬を得るまでに無限の行動を要する. 本手法は, 各状態で最悪の状態遷移をした場合に最も少ない行動数で報酬を得ることが期待できるルールを学習する手法である.

§2 罰状態遷移確率推定〜罰回避政策の中に合理的政策が存在しない場合〜

一般には, 図 7 の合理的政策改善手続きを利用しても, マークが到達しない状態が存在する可能性がある. この場合, そのような状態から報酬を得るためには, いくつかの罰ルールを選択する必要がある. ここでは, 合理的政策改善手続きの中に罰ルール間の順序付けを行う機構を組み込むことを考える.

罰ルールとは, ある確率で罰状態に遷移する可能性のあるルールであり, 罰状態とは, 罰ルールもしくは非合理的ルールしか選択できない状態である. そのため罰ルールの中でも, 極力, 罰状態に遷移しにくいルールを選ぶことが重要となる. 本論文では, 罰状態は, すべて等しく避けるべき状態であると仮定し, 各罰ルールの罰状態への

遷移確率を推定し, その値が最小である罰ルールを優先的に選択することを考える.

罰ルールの罰状態への遷移確率を**罰ルール度**と呼ぶ. 罰ルール度を推定するひとつの方法として, 次のような信頼区間の上界値を利用することを考える.

$$P_{ub} = 1 - \frac{\frac{x}{n} + \frac{Z_\gamma^2}{2n} + \frac{Z_\gamma}{\sqrt{n}}\sqrt{\frac{x}{n}(1-\frac{x}{n}) + \frac{Z_\gamma^2}{4n}}}{1 + \frac{Z_\gamma^2}{n}} \quad (1)$$

ここで n はそのルールの選択回数, x はそのルールによりその時点で判明している罰状態でない状態に遷移した回数, Z_γ は信頼度 95%のとき 1.96, 99%のとき 2.58 となる正規分布の両側パーセント点である. この式は, 罰が二項分布にしたがって与えられ, かつ,n が十分大きい場合に, 統計理論の区間推定の考えから導かれる [Kaelbling 91]. また P_{ub} の初期値は 0 とする. P_{ub} 値が 0 の場合のみ, そのルールは罰ルールではない. P_{ub} 値が 1 に近付くほど, そのルールはより罰状態に遷移しやすいルールになる.

信頼区間を利用する方法は正確であるが, 状態数の 2 乗のオーダーのメモリを要する. メモリを節約するために, 上で述べた罰ルール度の推定を数エピソードごとに行い, 過去のエピソードにおける情報ほど以下のような式で割り引いて考慮することを考える.

$$P_0 + \gamma P_1 + \gamma^2 P_2 + \gamma^3 P_3 + \cdots + \gamma^n P_n \quad (2)$$

ここで P_i はその時点で判定に利用した数エピソードにおける罰ルール度, γ は割り引き率で $0 < \gamma < 1$. さらに (2) 式の上限値を 1 とするために (2) 式全体に $\frac{1-\gamma^{n+1}}{1-\gamma}$ を乗ずる. いま S_{n-1} という推定値が得られている状況下で, P_0 という推定値が得られた場合, 新しい推定値 S_n は以下の式で与えられる.

$$S_n = \frac{\gamma - \gamma^{n+1}}{1-\gamma^{n+1}} S_{n-1} + \frac{1-\gamma}{1-\gamma^{n+1}} P_0 \quad (3)$$

現状態で, 罰ルールを選択する必要が生じた場合, 上記いずれかの方法で, 現状態で選択可能な各罰ルールの罰ルール度を推定し, 最も罰状態に遷移する可能性の低いルールを選択すればよい. その結果, 罰状態に遷移する確率が最も低い合理的政策の形成が保証される.

3·4 罰回避政策形成アルゴリズム

3.2 および 3.3 節で述べた手法を統合した**罰回避政策形成アルゴリズム**を図 10 に示す.

まず, 非合理的ルール判定用の 1 次記憶領域ならびに合理的政策記憶用の 2 次記憶領域を用意し, それらの内容を初期化する. 最初は, すべてのルールを非合理的ルールとみなし, 非合理的ルール集合に登録する. さらに, 罰ルール集合を初期化し, 罰ルール度を推定するために必要な記憶である (1) 式における n と x を初期化する.

初期化終了後, 状態 x_i を知覚し, **環境探査戦略**による行動 a_i を出力し, その行動を 1 次記憶領域の x_i 上に上書きする (図 11). 環境探査戦略としては, **ランダム探査**, **最少選択優先法** [宮崎 95], **k-確実探査法** [宮崎 95], 2 次記憶領域に記述されている行動をそのまま出力する **2 次記憶領域に基づく行動選択法**などが考えられる.

```
procedure 罰回避政策形成アルゴリズム
begin
  1次および2次記憶領域の内容を初期化する.
  すべてのルールを非合理的ルール集合に登録する.
  罰ルール集合および罰ルール度を推定するために
  必要な記憶を初期化する.

  状態xiを知覚する.
  do
    環境探査戦略による行動aiを出力する.
    aiを1次記憶領域のxi上に上書きする.

    if 報酬を得た then 1次記憶領域に存在する
       状態-行動対を非合理的ルール集合から除外する.
    if 報酬または罰を得た then call (罰ルール判定手続き).

     罰ルールおよび非合理的ルールを除外したルール
     集合に対し, call (合理的政策改善手続き).
     if 政策が未決定な状態が存在する then その状態で
       選択可能な各ルールの罰ルール度を計算し,
       その値の最も低いルールを政策に登録する.
     得られた政策を2次記憶にコピーする.

    遷移後の感覚入力xiを知覚する.
    罰ルール度を推定するための記憶を更新する.
  while
end.
```

図 10 罰回避政策形成アルゴリズム.

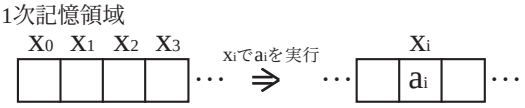

図 11 1 次記憶領域の更新方法.

行動を実行した結果, 報酬を得たならば, その時点で 1 次記憶領域に存在する状態-行動対は合理的ルールである. したがって, それらのルールを非合理的ルール集合から除外した後, 罰ルール判定手続きを起動し, 罰ルールを発見する. また, 罰を得るたびに, 新たな罰ルールが発生する可能性があるので, 罰を得た場合にも, 罰ルール判定手続きを起動する.

その後, 罰ルールおよび非合理的ルールを除外したルール集合に対し, 合理的政策改善手続きを適用する. その結果, もし政策が未決定な状態が存在するならば, その状態で選択可能な各ルールの罰ルール度を計算し, 最も値の低いルールを政策に登録する. 最後に, 得られた政策を 2 次記憶にコピーする.

さらなる政策の改善を目指すために, 遷移後の状態を知覚し, 罰ルール度を推定するための記憶を更新する. 以上を繰り返すことで,2 次記憶領域に罰を回避する合理的政策が形成される.

3·5 罰回避政策形成アルゴリズムの特徴

本アルゴリズムでは, 報酬と罰を別々に取り扱い, 合理的政策の獲得を大前提としている. そのため, どのように報酬および罰の値が設定されていたとしても, 期待獲得報酬量がゼロである政策が形成されてしまう事態をつねに回避することができる.

政策の形成に PIA などの動的計画法に基づく手法を利用しない, いわゆる Bellman-free[Sutton 98] な手法である. コストのかかる Backup 処理 [Sutton 98] が不要なため, PIA などが適用困難な大規模な状態空間を有する問題へも適用することができる.

提案手法は, 罰回避政策中の合理的政策の有無に応じた適応的なアルゴリズムである. 罰を完全に回避可能な政策が存在する場合には, それを発見し, そのような政策が存在しない場合には, 極力, 罰を回避する合理的政策を形成することができる.

提案手法は, 1 種類の報酬,1 種類の罰しか扱っていない. 複数種類の罰が存在し, それらの間の優先順位が既知な場合には, それぞれの罰の種類ごとに独立に罰ルールを判定し, 行動選択時に, 各罰の優先順位を考慮すればよい. 多種類の報酬に関しては今後の課題である.

4. 罰回避政策形成アルゴリズムの評価

4·1 Q-learning との比較

2 章で用いた, 環境 E2(図 4) で (R1,R2,P) を変化させ, Q-learning (QL) と罰回避政策形成アルゴリズム (Penalty Avoiding Rational Policy Making : $PARP$) を比較した. 図 12a) に $R1 = 50, R2 = 100$ とした場合, 図 12b) に $R1 = 0, R2 = 100$ とした場合の結果をそれぞれ示す. 各図の横軸は P の値, 縦軸は乱数の種を変えて行なった 100 回の実験における合理的政策の獲得率である. 特に,$R1 = 50, R2 = 100$ とした場合には, 罰を完全に回避する合理的政策が存在するので, 縦軸にはその政策の獲得率を示した. QL の学習率は 0.05, 割引率は 0.9, 行動選択には Q-値に基づくルーレット選択を用いた. 罰回避政策形成アルゴリズムでは, 環境探査戦略としてランダム選択を用いた.

$R1 = 50, R2 = 100$ とした場合, $PARP$ では, 任意の $P(< 0)$ に対し, S0 で a0 を選択する罰を完全に回避する合理的政策が形成された. 一方,QL では, P の値によっては, 必ずしもそのような政策が得られるとは限らない.P が小さい, すなわち罰が大きい場合には, 罰を完全に回避する合理的政策を獲得しているが, 罰が小さい場合には, 罰を完全に回避する合理的政策が存在するにも関わらず, 罰を得る可能性がある政策を得ている.

次に,$R1 = 0, R2 = 100$ の場合, $PARP$ では, 任意の $P(< 0)$ に対し, S0 で a1 を選択する合理的政策がつねに形成された. 一方,QL では, 罰が大きい場合, 必ずしもそのような政策が得られるとは限らない. P が大きい, すなわち罰が小さい場合には, 合理的政策を獲得しているが, 罰が大きい場合には, S0 で a0 を選択する報酬の得られない政策が学習される. QL の場合,P の変化に対する合理的政策の獲得率は, 図 12 の a) と b) とで全く正反対の結果となっている.

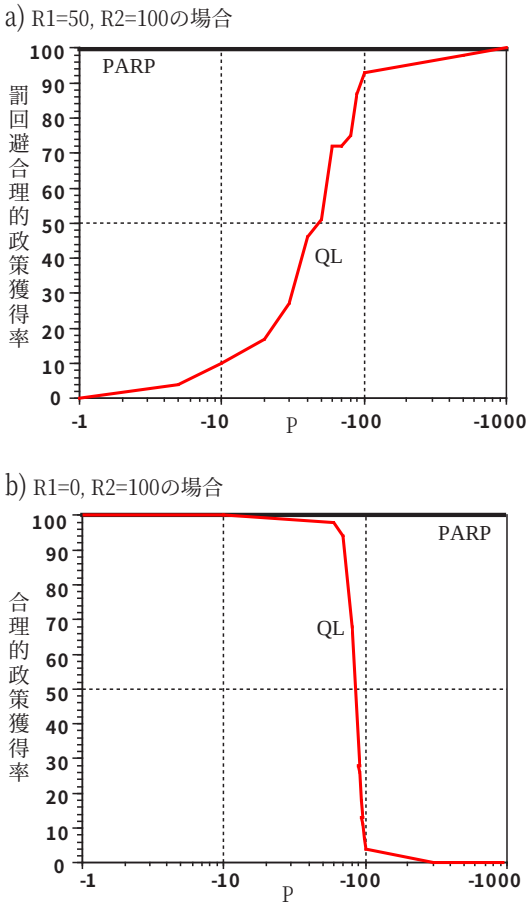

図 12 環境 E2 における Q-learning(QL) および罰回避政策形成アルゴリズム ($PARP$) の学習結果. a)R1=50,R2=100 の場合. b)R1=0,R2=100 の場合.

このように QL では, 期待する学習を実現する報酬および罰の設計は, 一般には容易ではない. それに対し,$PARP$ は, 任意の報酬および罰の値の下で, つねに合理的政策が形成されることが確認された.

4·2 tick-tack-toe への適用

§ 1 実験問題

罰回避政策形成アルゴリズムの性能を評価するために, ここでは 3 目並べの一種である tick-tack-toe に適用することを考える. tick-tack-toe とは, 3×3 の格子上の盤面に, 2 人のプレイヤーが交互に「○」と「×」を打っていくゲームである. 縦, 横, 斜めのいずれかに自分の記号を 3 つ並べた方が勝者となる. 勝者が決まらず盤面全体が埋めつくされれた場合には, 引き分けとなる. 状態数は 3^9 である.

本論文では, 罰回避政策形成アルゴリズムを図 13 に示

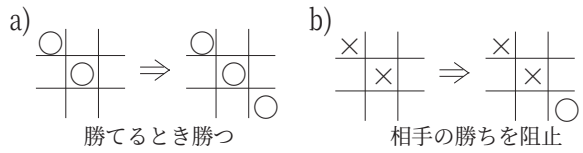

図 13 相手の知識.

表 1 報酬および罰の設定.

	勝ち	引き分け	負け
plan1	報酬	罰	罰
plan2	報酬	報酬	罰

すような知識を持った相手と戦わせることを考える．ここで知識とは，「自分が勝てるときは勝つ」(図 13a) が，そうでなければ,「相手の勝ちを阻止する」(図 13b) というものである．これらの知識は学習器には与えていない．また, 対戦相手は上記の知識が利用不可能な場合には, つねにランダムな手を選択するものとする．

一般に強化学習の実行結果は，報酬 および罰の設計方法に大きく依存する．特にゲーム問題では，引き分け時に罰を与えるか報酬 を与えるかで， 表 1 に示すような２通り考えられる．tick-tack-toe には必勝パターンは存在せず, 両者がエキスパートの場合, 引き分けになる．したがって,plan1 のように引き分け時に罰を与えると, 積極的に勝つための手を学習しようとする．一方,plan2 のように引き分け時に報酬を与えれば, 負けないための政策を試合全般にわたって学習することが期待できる．

もし必勝パターンが存在するゲームならば, plan1 のみでそれが学習可能であり,plan2 は全く必要ない．しかし先にも述べたように tick-tack-toe には必勝パターンが存在しないため, 負けないために, まず第一に, plan2 での罰ルールを除いた空間内のみで手を選択することが重要となる．その下で, plan1 での罰回避政策を選択することが可能ならば, その手が必勝手と言える．

以下では, 環境探査戦略として, 2 次記憶に基づく行動選択を行う罰回避政策形成アルゴリズムを考える．その下で, 必勝手が存在しない場合, 任意に決定した手をつねに打ち続ける方法を $PARP^-$, plan1 で罰状態に最も遷移しにくい手を 3.3.§ 2 における (1) 式で計算し選択する方法を $PARP$ と呼び, 比較実験を行なう．

§ 2 実験結果

各手法における, 乱数の種を変えて行った 100 回の実験での勝率および敗率を測定した．学習器が先手の場合を図 14, 後手の場合を図 15 に示す．tick-tack-toe は, 後手が不利なゲームなので, 後手の場合には, 引き分け率も同時に示した．横軸が試合数, 縦軸が勝敗 (引き分け) 率,S.D. は標準偏差を意味する．

tick-tack-toe には, 負けない手が存在するため, 先手, 後手ともにそのような手が必ず学習されている．勝率に関しては, 先手, 後手ともに, $PARP^-$ より $PARP$ の方

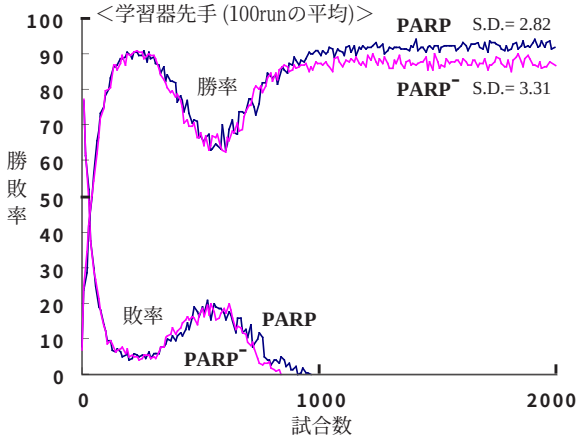

図 14 学習器が先手の場合の結果.

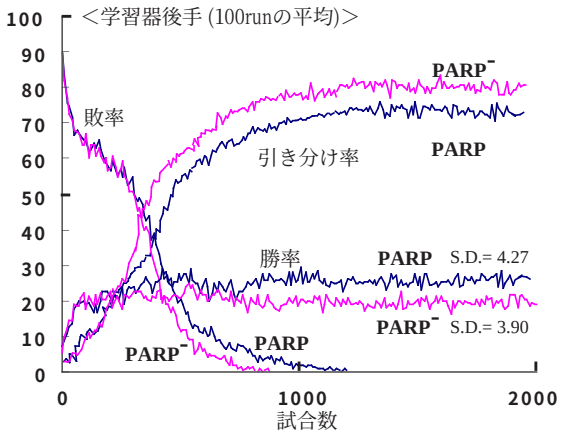

図 15 学習器が後手の場合の結果.

がよい．このことは，3.3.§ 2 で述べた方法が有効に機能していることを意味する．

tick-tack-toe の場合, 盤面が 3 × 3 と非常に小さいので, 一般に, 後手より先手の方が有利であり, 勝率の違いにそれがみてとれる．さらに, 先手の場合, 一度勝率が下がり, その後, 再び上昇している．これは, 一見よいと思った手が, 実は, 必勝手ではなかったことを意味する．相手の行動選択にはランダム要素があるため, 相手の選択可能な手のうちのごく一部のみが, 必勝手を妨げる手の場合, このような現象が生じる．tick-tack-toe は小さな問題であるが, $PARP$ の $PARP^-$ に対する有効性を確認することができた.

5. お わ り に

強化学習の工学的応用を考えた場合, 必ずしも最適政策の獲得が重視されるとは限らない．特に, 勝ち負けが存在するゲーム問題では, 負けない, すなわち罰を回避する政策が重視される．

本論文では, 罰を回避するための強化学習手法として, **罰回避政策形成アルゴリズム**を提案した．そこでは, 任意

の報酬および罰の値の下で, 罰を完全に回避する合理的政策が存在する場合にはそれを発見し, そのような政策が存在しない場合には, 極力, 罰を回避する合理的政策の獲得を実現している.

Q-learning との比較を行った後に, 3 目並べの一種である tick-tack-toe を取り上げ, 提案手法の有効性を確認した.

本手法をもとにオセロゲームへ適用した事例に [坪井 2000] がある. そこでは, 弱い罰としての**準罰**の導入や状態削減のための工夫など広範囲にわたる大幅でかつ重要な改良が施されており, 詳細は追って別の論文で報告する予定である.

さらに今後は, 提案手法と Profit Sharing[宮崎 94] とを組み合わせることで罰を陽に扱う Profit Sharing の確立や, 複数種類の報酬および罰が存在する場合への拡張を行いたい. その後は, 不完全知覚環境下 [宮崎 99a] やマルチエージェントシステム [宮崎 99b] などのより困難な問題クラスへ本手法を発展させていく予定である.

◇ 参 考 文 献 ◇

[Kaelbling 91] Kaelbling, L. P.: An Adaptable Mobile Robot, *Proceedings of the 1st European Conference on Artificial Life*, pp. 41–47 (1991).

[Papadimitriou 87] Papadimitriou, C. H. and Tsitsiklis, J. N.: The Complexity of Markov Decision Processes, *Mathematics of Operations Research*, pp. 441–450 (1987).

[Sutton 98] Sutton, R. S. & Barto, A.: Reinforcement Learning: An Introduction, *A Bradford Book*, The MIT Press (1998).

[坪井 2000] 坪井創吾, 宮崎和光, 小林重信: 罰回避政策の形成とゲーム問題への応用, 第 27 回知能システムシンポジウム資料, pp.117-122(2000).

[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.124-133 (1995).

[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, 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.148-156(1999).

[ワグナー 78] ワグナー (高橋 幸雄, 森 雅夫, 山田 尭 訳). **「オペレーションズ・リサーチ入門** *5*=**確率的計画法」**, 培風館, (1978).

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

〔担当委員：阿久津達也〕

2000 年 4 月 3 日 受理

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, 日本機械学会各会員.

坪井 創吾

1998 年東京工業大学生命理工学部生物工学科卒業. 2000 年同大学大学院総合理工学研究科知能システム科学専攻博士前期課程修了. 同年, 東芝（株）に入社. 現在, 同社研究開発センターに所属. 知識情報共有システムおよびコミュニティ支援技術に従事.

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.