緩い対称性推論を用いた強化学習アルゴリズム
Reinforcement Learning Algorithm using Loosely Symmetric Reasoning

甲野 佑†, 高橋 達二‡
Yu Kohno, Tatsuji Takahashi

† 東京電機大学大学院先端科学技術研究科, ‡ 東京電機大学理工学部
Tokyo Denki University, Tokyo Denki University
yu.kohno.02@gmail.com, tatsujit@mail.dendai.ac.jp

Abstract

We applied the *LS* (loosely symmetric) model proposed by Shinohara to reinforcement learning. For that purpose, we analyzed some distinctive properties of LS in decision-making including self-organization of policy, resulting from symmetric inference by the model. Based on the analysis, we constructed a reinforcement-learning algorithm using LS as the value function. We showed the new algorithm 's efficacy in Acrobot task that has nonlinear dynamics. This study prepares future extensions of the model.

Keywords — Symmetric reasoning, Decision-making, Reinforcement learning

1. はじめに

現実に生きる我々は，観測，記憶，思考，行動に量的な制限を常に受け続けている．予測の正確さと判断の早さはトレードオフの関係にあり，全ての事柄を正確に知り，熟慮した上で，かつ素早く答えを導く事は出来ない．このトレードオフは何も知らないエージェントが未知の環境に対して学習をして行く際に最も顕著に現れる．自ら主体的に行動しなければ知識は得られない．かといって，限られた知識から何を選択していくべきか決定するのは困難である．本研究ではこのような困難を抱える学習課題(強化学習[1])に対して対称性推論を用いたアプローチを行った．ここで定義する推論とは，与えられた前提から結論を導く事や，得られた情報から結果を予測する事を指す．そして対称性推論とは人の直感的推論能力に深く関わっているとされる推論形式である．“p ならば q” が真という情報を得た時，“q ならば p” も真であると考えてしまう人の直感的傾向を対称性バイアスと言い，対称性推論にはそのようなバイアスによる非論理的な傾向が見られる．対称性推論は規範的論理学に照らし合わせると誤った推論であるが，現実の環境ではむしろ有用であるとされる[2].

意思決定課題において対称性推論が如何なる影響を与えるのかは，前述のトレードオフを抱える最も基礎的な課題であり，最も単純な強化学習課題でもある2本腕バンディット問題を例に説明されている．篠原らは緩い対称性推論のモデルである篠原モデル(緩い対称モデル，*LS*)を考案し，前述の2本腕バンディット問題に対してシミュレーションを行い，対称性推論が意思決定課題に対して良い成績を持つ事を示した[3]．本研究ではこの *LS* を強化学習アルゴリズムへ応用する事を目的としている．しかし，現在の形式では *LS* モデルを一般的な強化学習課題に対して応用する事が出来ない．そこで以下ではまず *LS* の性質を再考察し，どのような形式で強化学習への一般化を行うべきか議論する．

2. Loosely Symmetric Model

LS は人間の因果帰納と高い相関を持つモデルであり，$LS(E|C)$であれば，原因候補事象 C から結果事象 E が起こる事に対する信念の度合いを計算する(式1).

$$LS(E|C) = \frac{P(C,E)+S_p}{P(C,E)+S_p+P(C,\bar{E})+S_n} \quad (1)$$

$$\text{Positive bias}: S_p = P(\bar{E})P(C|\bar{E})P(\bar{C}|\bar{E}) \quad (2)$$

$$\text{Negative bias}: S_n = P(E)P(C|E)P(\bar{C}|E) \quad (3)$$

また，Takahashi et al. [4] によって，*LS* の性質と視覚における地と図の関係との類似が指摘されており，多義的に人の認知的性質を含むことがわかっている．人に対する認知実験の結果との統計的な比較もなされており，例えば人の因果帰納実験の結果に対して高い相関を得ている[5]．人に対するバンディット問題の十分な行動実験がまだ行われていない現状では “高い記述能力” と “高いパフォーマンス” という基準の違いは有るものの，因果帰納と意思決定という異なる分野において有意な結果を残しているモデルは *LS* のみである[6]．故に，対称性とそれに付随する諸性質の理解を深めるに当たって非常に重要なモデルになり得ると考えられて来た．

2.1 対称性推論と相対評価

LS は対称性推論という人の認知的性質を参考にして考案された[3]. 対称性推論とは論理的な推論において条件文 "C ならば E である $(C \rightarrow E)$" が真あるという前提がある時, 逆命題である "E ならば C である $(E \rightarrow C)$" も真であると, 与えられた条件文を双条件的に捉えてしまう人間の認知傾向である. これは誤りであるが, 人間の生活環境では双条件的に解釈する事が正しい場合も多い[7].

更に, 人間は裏命題である "C でないなら E でもない $(\bar{C} \rightarrow \bar{E})$" も真であると考える傾向も知られている. これは対称性推論の推論形式の一つであり, 相互排他性と呼ばれる. 例えば, ある計画が失敗した場合, やってもいない他の計画への期待(評価)が上昇する事等が挙げられる. このように, ある手段に対する試行結果が直接的な関係の無い(あるいは関係あるかどうか解らない)他の手段の評価に影響する事は, 規範的な論理学から導出されない. しかし, これは我々人間が普段よく行ってしまう評価形式である. このような評価形式は相対評価と呼ばれ, ある手段が上手くいけば其れに執着し, 上手くいかなければ他の手段を試すよう促す効果を生む. つまり, 対称性推論では論理的に関係があるとは限らない事象を関連づける事で, 客観的, 絶対的な評価ではなく, 主観的, 相対的な評価を可能としている.

意思決定課題で LS が優れた結果を残しているのは, LS が対称性推論により, この相対評価を行う事に理由がある[3]. しかし, LS の持つ優れた性質はそれだけではない. 以降では LS のどのような性質が具体的に学習に有効であるのか考察する.

2.2 LS を学習に用いる利点

LS はある事象 C の発生(知識)に対する観測頻度から, その知識が正しいか間違っているかの不確かさを価値の評価に加える事が出来る. ここで定義する価値とは, 目的とする特定の結果(報酬獲得)をどれだけ達成できるかを意味する. これは学習過程において, 未知の環境から試行と観測で偶然とも必然とも解らない知識を学習して行くために重要な性質であると言える. しかも, これは相対的であり, 最も観測した知識の評価はより客観的, 統計的な指標(期待値等)に近づいて行く.

$$\lim_{P(C)\rightarrow 1.0} LS(E|C) \approx P(E|C) \quad (4)$$

$$\lim_{P(C)\rightarrow 0.0} LS(E|C) \approx 0.5 \quad (5)$$

数式の上では, 事象の相対的知識量 $P(C)$ の増加, 減少に対する LS の極限(式4, 5)から, 曖昧さの評価が端的に表される. $P(C) \rightarrow 0.0$, 即ち事象 C について殆ど知らない場合, 他の条件がなんであれ, LS の評価は 0.5 に固定される. この 0.5 という値は事象 C から事象 E が発生するか否かの確率としては, 最も曖昧なものである. 逆に $P(C) \rightarrow 1.0$, 即ち事象 C についてよく知っている場合, LS の評価は事象 C から事象 E が発生する事に対して, 最も客観的な指標である条件付き確率 $P(E|C)$ になる. 後述する強化学習[1]のような, 自ら主体的に行動し, 情報を獲得して行かなければならない学習において, この曖昧さを評価して意思決定して行く性質は特に重要になると考えられる.

2.3 環境適応と自己組織化

LS の極限である式4, 5から, $P(E|C)$ の値が 0.5 である場合を境にLSの評価が逆転する. この評価の境となる点を参照点と呼ぶ. LS の最も興味深い性質は, このような参照点を持つ事で, 単純な式ながら不完全な観測情報から環境の状態をマッピングし, ポリシーを多様に変化させる事にある. これは一つの価値関数が, 複数のポリシーを自ら選択するという点で, 一種の自己組織化を行っているのだと考えられる.

最も単純な強化学習の一種に, 2本腕バンディット問題が存在する. 2本腕バンディット問題とは, 事前に報酬の獲得確率が知らされていない2種の選択肢から, どちらを試行するか決定し, 確率的に得られる報酬を最大化させるよう学習する課題である. より報酬獲得確率が高い選択肢のみを試行すれば, 当然最終的に得られる確率は最大化される. しかし事前に知識も無い状態では, どちらの選択肢の報酬獲得確率が高いか知る事は出来ない. 必ず, それらの選択しに対する試行錯誤, 探索を行わなければ報酬獲得に関する知識を得る事は出来ない. この点で探索と報酬獲得のジレンマを抱える課題の中でも, 報酬の最大化の困難さを最も端的に表す課題であると言える.

この課題において, LS はどの手段も報酬獲得確率が参照点である 0.5 未満の不毛な環境ではよく探索し, どの手段も報酬獲得確率が参照点以上の肥沃な環境では報酬を最大化させる傾向が示されている[9]. 即ち LS には『少なくともいずれか一つの手段は 0.5 以上の報酬確率を持つであろう』というポリシーを持ち, 報酬確率 0.5 以上の手段

を求めて探索を行う傾向が存在する．しかもこの方策の変化は緩く，連続的に行われている(図1)．単一のモデルでこのように環境から得た不完全な情報を用い，連続的で多様なポリシーの切り替えられるのは *LS* のみである．

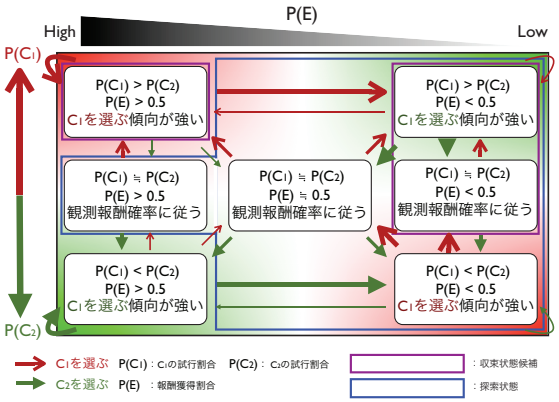

図 1 *LS* の緩い方策状態遷移

LS の方策状態の遷移はこの参照点と観測値との間で緩く行われている．故に *LS*は飢えているからと言って極端に探索を行うような事はせず，空腹の度合いと現在得られている情報から，未知の手段の探索と，既知の手段に対する報酬獲得の配分を自己で適宜変化させる事が出来る．空腹の度合いはロボットで言えばバッテリーの残量等に当たる．事前に把握しきれない複雑な構造を持つ，あるいは時間とともに変化する非定常な環境では，このような“飢え”と環境の相互作用によって探索的選択の度合いを自己組織化的に決定する性質が，人工知能に欠かせない能力になるだろう．

また，*LS* は2本腕バンディット問題のみならず，選択肢を N 種に一般化した課題である N 本腕バンディット問題においても優れた成績を残している[8]．しかし，既存の *LS* は飽くまで客観的な確率を変数として，主観的な確率の値を算出するモデルに過ぎず，強化学習に使われるQ値等の無限の値域を持つ価値関数には対応できていなかった．以降では，強化学習の基本的な性質に触れながら，具体的に強化学習への *LS* の一般化を論じる．

3. 強化学習への応用

本研究の目的は，*LS* の持つ性質を活かし，新たに強化学習へ応用可能なアルゴリズムの構築である．強化学習とは，環境と学習エージェントの相互作用から最適な方策を学習する機械学習の一種である．エージェントは自身と環境の状態 s を把握し，その状態 s において可能な行動群 A の中から実際に取る行動 a を方策 π に基づき選択し，次の状態 s' に推移．報酬 r を得る．これらの情報

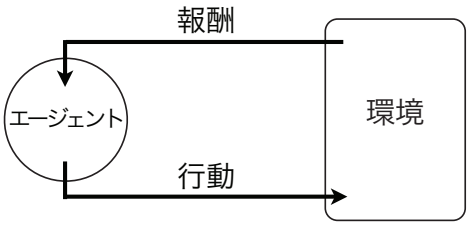

図 2 強化学習の概要

から状態行動対$(s,\ a)$の価値，そして行動方策を学習する．強化学習と他の学習形式(教師あり／なし学習)の最大の相違点は，エージェント自身が主体的に環境から情報を探索しなければならない事である．報酬を得る事のみを優先すれば探索は進まず，かといって情報探索に時間を費やせば最終的に得られる報酬は少なくなってしまう．他にも強化学習の抱える問題は連続状態，連続行動に対する価値関数の近似等，様々存在する，しかし，このような収穫と探索のジレンマを抱えている以上，少なくともその問題においては *LS* を強化学習に用いる事は有用であると思われる．

強化学習に対する *LS* の応用は清水ら [11]や，Uragami [12]によってアプローチされており，相対的に優れた結果を得ている．しかしそれらは情報の更新に部分的に用いられたのみで，実際の行動選択においての参照点による自己組織化的性質については重きが置かれていなかった．以下では *LS* を行動選択の方策 π として扱う事を目的とし，具体的には $TD(\lambda)$ 学習の一種である $Sarsa(\lambda)$ に *LS* を応用するため，幾つかの変更を行った．

3.1 $Sarsa(\lambda)$

$Sarsa(\lambda)$ は $TD(\lambda)$ 学習の一つである．まず *Sarsa*の最も単純な形式である $Sarsa(0)$ について説明する．$Sarsa(0)$ はマルコフ過程を前提とし，Q値と呼ばれる状態行動対 (s,a) に対する価値関数 $Q(s,a)$ を以下の Bellman 方程式に従い学習する．

$$Q(s_i,a_j) \leftarrow Q(s_i,a_j) + \alpha[r+\gamma Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)] \quad (6)$$

特に以下の式7を *TD* 差分と呼び，現在の価値と得られた情報の差異を意味している．この *TD* 差分を一つ前の状態にバックアップする事で，オンラインに学習する事が出来る．

$$[r+\gamma Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)] \quad (7)$$

ここで，α は得られた情報をどれほど学習するかを決める学習率$(0.0 \leqq \gamma \leqq 1.0)$，$\gamma$ は未来に得られる価値をどれだけ減衰させるかという割引率である$(0.0 \leqq \gamma \leqq 1.0)$．割引率は低い程即時報酬，即ちすぐに手に入る報酬に価値を見いだすようになり，高ければ将来手に入る報酬を高く見積もるようになる．

$Sarsa$ は方策オン型の TD 学習とされる．それは実際の行動を選択する際に方策 π を使うのみでなく，バックアップに使う行動 a_{t+1} を選択する際にも方策 π を用いる事を意味する．方策オフ型 TD 学習では Q 学習が知られている．Q 学習の $Sarsa$ との違いはバックアップに使う行動には常に価値価数が最大の物を選択する(greedy)事のみである．方策オフ型学習の Q 学習ではなく方策オン型学習の $Sarsa$ を LS エージェントに用いる理由は，連続状態に対する価値関数の線形関数近似に対して収束する事が知られているためである[1]．$Sarsa(0)$ は一つ前の状態に関数近似するのみで，大きな状態空間に対してオンラインに学習する事が困難である．これを解決するために，$TD(\lambda)$ 学習への拡張が必要になる．以下の式は Bellman 方程式を $Sarsa(\lambda)$ に拡張した物である．

$$Q(s_i,a_j) \leftarrow Q(s_i,a_j) + \alpha[r+\gamma Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)]e(s_i,a_j) \tag{8}$$

$e(s_i,a_j)$ は適格度トレースと呼ばれ，ある状態行動対 (s_i,a_j) が，現在の状態行動対 (s_t,a_t) に至るまでにどの程度影響を及ぼしたかの指標である．$Sarsa(0)$ では一つ前の価値関数 $Q(s_t,a_t)$ しか更新しなかったが，$Sarsa(\lambda)$ では全ての $Q(s_i,a_j)$ に対し，適格度トレースに従って TD 差分の加算が行われる．全ての状態行動対に対する適格トレースは，1 ステップ毎に以下の式によって減衰する．

$$e(s_i,a_j) = \lambda e(s_i,a_j) \tag{9}$$

訪問する毎に適格度トレースは上昇する．本研究では，初回訪問型と呼ばれる更新法に従い訪問した状態行動対の適格度トレース $e(s_t,a_t)$ に対し，以下の式を用いた．

$$e(s_t,a_t) = 1.0 \tag{10}$$

方策オフ型の $TD(\lambda)$ 学習である $Q(\lambda)$ 学習の更新法には幾つかの形式があり，万能性や唯一性(最適性)は保証されていないが，方策オフ型の $TD(\lambda)$ 学習である $Sarsa(\lambda)$ 学習は前述のように $Sarsa(0)$ 学習の TD 誤差を適格度トレースとの積に変更するだけで実装でき，方策 π に従う価値関数へ収束する事が保証されている．これも本研究において $Sarsa(\lambda)$ を用いた理由である．

3.2 タイルコーディング

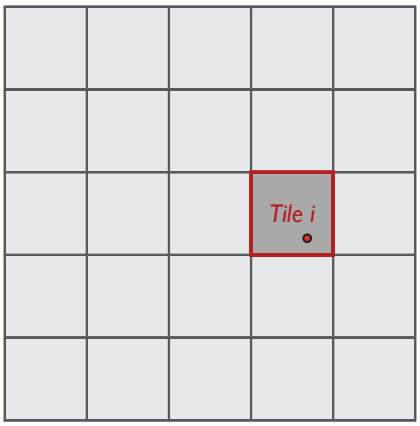

図 3 タイリング

前述で触れたように，強化学習は連続状態に対する関数近似を如何なる形で行うかという問題を抱えている．離散状態においても，次元数が増える程に状態数が指数的に増えてしまい，学習が上手く進まないという問題がある(次元の呪い)[1].

そこで，本研究ではタイルコーディングを使い，連続的な状態変数や，状態の次元が多い場合の離散化を行った．タイルコーディングではタイリングという概念を扱う．タイリングとは図(3)のように状態変数を連続状態から離散的に荒く分割する事である．しかし，分割数が少なければ正確な学習は行えず，多くすれば前述の次元の呪いのために状態数は指数的に増加する．その回避のため，図(4)のようにタイルコーディングでは複数のタイリングを用いる事で複雑な状態を表現した．各タイリングは座標をずらしており，ある座標の状態はタイリング数 c の長さを持つベクトルで表現される

$$\vec{s_i} = [s_{i1}, s_{i2}, s_{i3}, ..., s_{ic}] \tag{11}$$

$Sarsa(\lambda)$ の更新式は以下のように変更される．また，状態 $\vec{s_i}$ の価値はベクトルの要素の価値の和で表される．次の行動は，価値関数 $Q(\vec{s_i},a_j)$ を比較して方策 π に従い決定される．

$$Q(s_{ik},a_j) \leftarrow Q(s_{ik},a_j) + \tag{12}$$

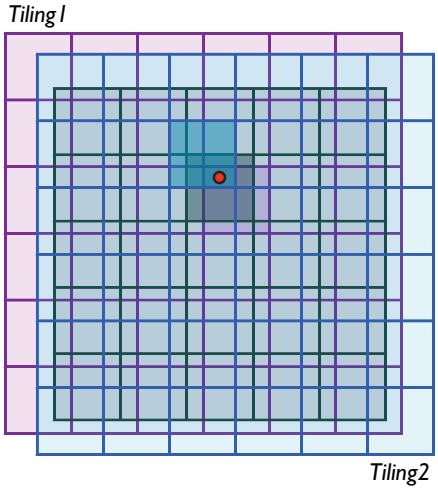

図 4 タイルコーディング

$$Q(\vec{s_i}, a_j) \leftarrow \sum_{k=1}^{c} Q(s_{ik}, a_j) \quad \frac{\alpha}{c}[r+\gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]e(s_{ik}, a_j) \tag{13}$$

タイリング毎にタイルの区切り方(離散化)は等しく無くても良く，またタイルの区切り方を工夫する事で特定の次元を無視する事や，特定の区間を細かく分割する事で詳細に価値関数を設定する事が出来る[1].

3.3 LSを用いた強化学習アルゴリズム

本研究では上述の性質を保ったまま LS の計算式をSarsa学習[1]に応用するため，学習で扱う変数を以下のように再定義した.

表 1 保持すべき情報

s	価値	試行頻度
a_1	Q_1	τ_1
a_2	Q_2	τ_2
$\vdots$		
a_n	Q_n	τ_n

τ_k：行動a_kを行った頻度
Q_k：行動a_kを行う価値

表1から，状態 s において行動 a_k を取った際の単位時間当たりの獲得報酬は Q/τ から以下の式になる.

$$R_k = \frac{Q_k}{\tau_k} \tag{14}$$

$$R_U = \frac{\sum Q_k}{\sum \tau_k} \tag{15}$$

この規準と単位が同じく (価値/試行度合い) になるよう，方策が変化する参照点 R_c を本研究では以下のように設定する．これはまだ試行錯誤中の値であり，より理論的に妥当な値がある可能性は否定しない.

$$R_c = \frac{\text{前エピソードで得た総報酬}}{\text{前エピソードの総試行数}} \tag{16}$$

これらの変数から状態 s において可能な任意の行動 a_k に対する価値関数 $LSRL$ を考案する(式17)．ここで a_H は最も観測した手段，即ち最も τ の高い手段 $\max_{\tau_k} a_k$ であり，同様に a_L には最も観測していない手段，即ち最も τ の低い手段として $\min_{\tau_k} a_k$ が選択される．これは LS のバイアス項，式2，3に習っている．また，実装に行う行動，バックアップに使う行動には最も高い $LSRL$ 値を持つ行動 a_k，つまり $\max_{LSRL(a_k)} a_k$ が選択される.

$$LSRL(a_k) = \frac{Q_k + 2R_c\frac{\tau_H\tau_L}{\tau_H+\tau_L} - \frac{Q_HQ_L}{Q_H+Q_L}}{\tau_k + \frac{\tau_H\tau_L}{\tau_H+\tau_L}} \tag{17}$$

エージェントが状態 s にあるとき，$t_H/(t_H+t_L) \rightarrow 1.0$ の極限において最も観測した行動 a_H，最も観測していない行動 a_L の LS 値は式18，19となる．これは通常の LS の極限式4，5に対応する.

$$\lim_{\frac{t_H}{t_H+t_L}\rightarrow 1.0} LSRL(a_H) \approx \frac{Q_H + 2R_ct_L - Q_L}{t_H + t_L} \approx \frac{Q_H}{t_H} \tag{18}$$

$$\lim_{\frac{t_H}{t_H+t_L}\rightarrow 1.0} LSRL(a_L) \approx \frac{Q_L + 2R_ct_L - Q_L}{t_L + t_L} = R_c \tag{19}$$

4. 強化学習シミュレーション

以下では $LSRL$ を実装したエージェントが，そうでないエージェントに比べてどのような成績の違いを持つか調べるため Acrobot 振り上げ課題で学習を行い結果を比較した．比較するエージェントには全て方策オン型学習 $Sarsa(\lambda)$ を用い，方策 π には．ϵ - greedy と softmax 方策を用いた[1]. ϵ - greedy 方策は ϵ の確率でランダムに手段を選択し，$1-\epsilon$の確率で greedy に，最も Q 値の高い手段を選ぶ方策である．ϵ が高い程多く探索を行うため，ϵ の初期値を 0.5 とし，ステップが経過する毎に ϵ を減衰させた.

初期値 : $\epsilon_0 = 0.5$

ステップが経過する毎に,

$$\epsilon_{t+1} \leftarrow 0.9\epsilon_t \tag{20}$$

softmax 法とは，Boltzmann 分布に従って Q 値から行動選択確率(式21)を算出し，その確率に従って行動を決定する手法である．ϵ - greedy 探索する場合，全ての手段に対して等しい確率で探索してしまう．それに比較して softmax 方策は Q 値の大きさによって選択確率を変更できるため，より効率的に探索する事が出来る．

$$\pi(s_t, a_k) = \frac{exp(Q(s_t, a_k)/T)}{\sum_{i=1}^{n} exp(Q(s_t, a_i)/T)} \tag{21}$$

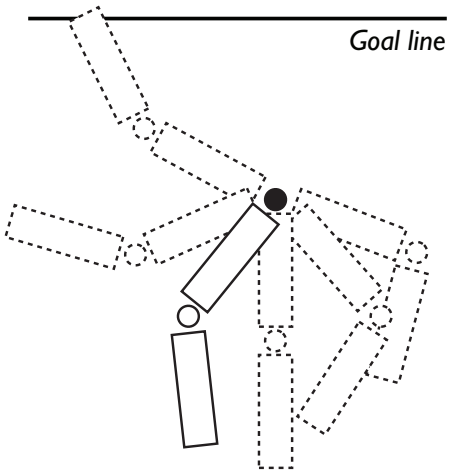

図 5 Acrobot 振り上げ課題の挙動

本シミュレーションで扱う Acrobot 振り上げ課題は，鉄棒のような回転可能な軸に繋がれたロボットが身体を揺らし，その終端が一定の高さ (Goal line) に達する事を目的とした課題である(図5)．課題の各設定は Sutton[10] に準じている．ロボットは下半身との間に一点，可動部を供えており，エージェントはその可動部に加えるトルク τ を学習する．ここで与えるトルク τ は正の向き $1.0Nm$ ($\tau = 1.0$)，負の向き $1.0Nm$ ($\tau = -1.0$)，トルクを与えない ($\tau = 0.0$) の3種とした ($\tau \in \{10.0, -10.0, 0.0\}$)．また，ロボットが繋がれた回転可能な軸に直接トルクを与える事は出来ない．

状態は，重力方向を $\theta = 0.0$ として回転軸に対する上半身の角度 θ_1 とその角速度 $\dot{\theta}_1$，上半身と下半身のなす角度 θ_2 とその角速度 $\dot{\theta}_2$ の4種変数によって表現される．それぞれの初期値は，$\theta_{10} = 0.0[rad]$，$\dot{\theta}_{10} = 0.0[rad/s]$，$\theta_{20} = 0.0[rad]$，$\dot{\theta}_{20} = 0.0[rad/s]$ である．上半身とか半身のなす角度 θ_2 は $-(3/4)\pi < \theta < (3/4)\pi[rad]$ の範囲に制限した．また，角速度 $\dot{\theta}_1$，$\dot{\theta}_2$ は $-4\pi < \dot{\theta} < 4\pi[rad/s]$ の範囲を取るよう設定した．回転軸に対する上半身の角度 θ_1 には制限を与えておらず，一つのエピソード中に何度でも回転可能である．

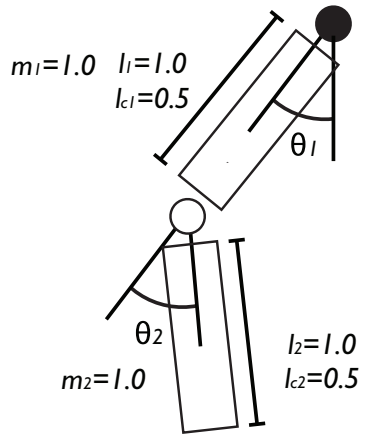

図 6 Acrobot のパラメータ

各ステップでは状態変数と与えられた力 F から以下の式によって回転軸と上半身との間の角加速度 $\ddot{\theta}_1$ と，上半身と下半身の間の角加速度 $\ddot{\theta}_2$ を計算した．使用される変数は台車の重さ $M = 1.0[kg]$，ポールの重さ $m = 1.0[kg]$，長さ $l = 0.5[m]$，重力加速度 $g = 9.8[m/s^2]$ である．

$$\ddot{\theta}_1 = -d_1^{-1}(d_2\ddot{\theta}_2 + \phi_1) \tag{22}$$

$$\ddot{\theta}_2 = (m_1 l_{c1}^2 + I_2 - \frac{d_2^2}{d_1})^{-1}(\tau + \frac{d_2}{d_1}\phi_1 - \phi_2) \tag{23}$$

$$d_1 = m_1 l_{c1}^2 + m_2(l_2^2 + l_{c2}^2 + 2l_1 l_{c2} cos\theta_2) + I_1 + I_2 \tag{24}$$

$$d_2 = m_2(l_{c2}^2 + l_1 l_{c2} cos\theta_2) + I_2 \tag{25}$$

$$\phi_1 = -m_2 l_1 l_{c2}\dot{\theta}_2^2 sin\theta_2 - 2m_2 l_1 l_{c2}\dot{\theta}_1\dot{\theta}_2 sin\theta_2 + (m_1 l_{c1} + m_2 l_1) g cos(\theta_1 - \pi/2) + \phi_2 \tag{26}$$

$$\phi_2 = m_2 l_{c2} g cos(\theta_1 + \theta_2 - \pi/2) \tag{27}$$

その後，以下に示すオイラー法によって差分を行い，状態変数である位置 X，速度 $\dot{X}$，角度 θ，角速度 $\dot{\theta}$ を決定する．差分に用いる時間増分 Δt は $\Delta t = 0.05[min]$ とした．

$$\dot{\theta}_1 \leftarrow \dot{\theta}_1 + \ddot{\theta}_1 \Delta t \tag{28}$$

$$\dot{\theta}_2 \leftarrow \dot{\theta}_2 + \ddot{\theta}_2 \Delta t \tag{29}$$

$$\theta_1 \leftarrow \theta_1 + \dot{\theta}_1 \Delta t \tag{30}$$

$$\theta_2 \leftarrow \theta_2 + \dot{\theta}_2 \Delta t \tag{31}$$

連続状態である状態変数はタイルコーディングを用いて，状態変数毎に6つに分割した．タイリングのパターンはシミュレーション毎にランダムに分割しなおし，パターンに依存しないよう設定した．一つのタイリングに付き状態数は $6\times6\times6\times6 = 1,296$ であり，一度のシミュレーションにつきタイリングを12パターン用いて計算を行った．環境からの報酬はステップの経過に伴い $r = -1.0$ を与え，Acrobot が Goal line を越えた際に $r = 0.0$ を与えた．Acrobot が Goal line に達するか，2,000 ステップ経過すると，1 エピソードが終了する．

エージェントは ϵ - greedy，softmax 方策を用いた2種類の $Sarsa(\lambda)$ と，本論文で新たに考案した $LSRL(\lambda)$ を用い，$\lambda = 0.0$ ($Sarsa(0.0)$ ϵ - greedy，$Sarsa(0.0)$ softmax，$LSRL(0.0)$) と，$\lambda = 0.9$ の場合($Sarsa(0.9)\epsilon$- greedy，$Sarsa(0.9)$ softmax，$LSRL(0.9)$)，合計6つのエージェントで各 500 エピソード，1,000回シミュレーションした．

4.1 結果および考察

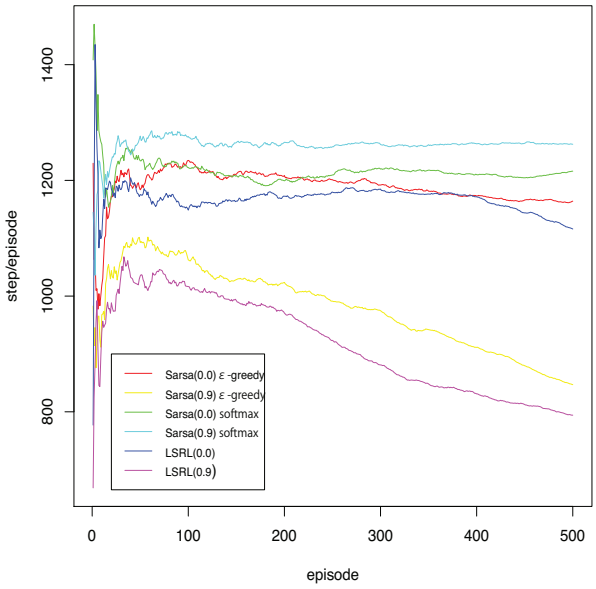

図 7 Acrobot 振り上げ課題において終端状態に到るまでのステップ数

図7は Acrobot 振り上げ課題における学習の過程を表している．横軸はエピソードの経過数であり，縦軸はそれまでのエピソードを通して経過したステップ数をエピソード数で平均した物である．

$\lambda = 0.0$ の時，ϵ - greedy と softmax を用いた $Sarsa$ が上手く学習出来ていないのに対して，$LSRL$ は比較的短いステップで終端状態に至った．$\lambda = 0.9$ の差異には ϵ - greedy よりもよく学習をし，早いステップ数で Goal line を越えている．Acrobot 振り上げ課題は状態変数が連続的であるだけでなく，非線形なダイナミクスを持つ難易度の高い課題として知られる[10]．この課題の荒い離散化(タイルコーディング)に対して $LSRL$ の学習の早さは，$LSRL$ が複雑な環境の情報を持つ課題に対してもロバストに対応出来る事を示している．今回は参照点 Rc (式16)を今までで最も良い結果に設定した．これは過去最も良い結果を招いた選択と比較し，同等か，それを越えるような選択を“常に”探索し続ける事を意味している．このように本シミュレーションにおいて非常にストイックな条件においても有意な結果を得たという事は，実際の物理現象に見合ったより最適な参照点 Rc を与える事で更に成績が向上するものと推測できる．

5. 総合考察

本研究では計算機上での効果的な学習・推論システムの構築のため，機械学習を根幹に認知心理学等の他分野の成果を応用し，人間を模倣するソフトコンピューティングの体系化に寄与することを試みた．具体的には，強化学習課題における対称性推論の有用性を示すため，緩い対称モデル，LS に着目し，強化学習への一般化を試みた．特に LS が対称性推論によって環境の複雑さを考慮し，環境毎にその挙動を変化させて課題に適応する性質や，変化の基準値である参照点とエージェントの状態を対応させる事で，更に柔軟な環境適応する性質を重視し，新たな強化学習モデルである $LSRL$ の構築を行った．また，そのシミュレーションから強化学習における対称性推論の有用性の一端を示唆するに至った．

諸研究により LS は人間の因果帰納実験や意思決定実験に対して高い記述性を示している[5]．それらの結果と本研究の分析から LS は偶然の影響，不確定な情報に適応する人間の認知能力を表現する良いトイモデルとなる可能性がある．現時点では LS を方策オン型学習に実装したに過ぎないが，人間の認知バイアスが学習を促すという側面を直接的に見る事が出来た．本研究の結果は LS の強化学習への応用において完全なものではない．しかし，強化学習一般に適用可能な形式を提示した事で対称性推論に関する研究に幅を持たす事にもつながり，Actor - critic 等のより生物的な学習への実装に見通しが付いた．

本研究の成果は，機械学習の形式で広い範囲に対応可能なモデル構築の一つの方針を示した．のみならず，対称性推論を現実環境の認知的な記号化とその学習システムにまで拡張し，現実で働く

人工知能に対する諸問題に対する具体的アプローチであると言える.

参考文献

[1] R. S. Sutton, A. G. Barto (2000), “強化学習”, 森北出版, (三上, 皆川 訳).

[2] M. Hattori and M. Oaksford (2007), “Adaptive non-interval heuristics for covariation detection in causal induction: Model comparison and rational analysis”, *Cognitive Science*, 31, 765–814.

[3] 篠原修二, 田口亮, 桂田浩一, 新田恒雄 (2007), “因果性に基づく信念形成モデルと N 本腕バンディット問題への適用”, 人工知能学会論文誌, Vol.22, No.1, 58–68.

[4] T. Takahashi, M. Nakano, S. Shinohara (2010), “Cognitive symmetry: Illogical but rational biases”, *Symmetry: Culture and Science*, Vol.21, No.1-3, 275–294.

[5] T. Takahashi, K. Oyo, S. Shinohara (2011), “A Loosely Symmetric Model of Cognition”, *Lecture Notes in Computer Science*, No. 5778, Springer, pp. 234–241.

[6] 大用庫智, 高橋達二 (2010), “因果帰納と意思決定を結ぶ緩い対称モデル”, 日本認知科学第*27*回大会, P3-34.

[7] T. Takahashi, M. Nakano, S. Shinohara (2010), “Cognitive symmetry: Illogical but rational biases”, *Symmetry: Culture and Science*, Vol. 21, No. 1-3, 275–294.

[8] 大用庫智, 甲野 佑, 高橋 達二 (2011), “非定常 N 本腕バンディット問題に対する人間の認知バイアスの適用”, 2011年度人工知能学会全国大会, *2011*年度人工知能学会全国大会*(*第*25*回*)*予稿集, 1G1-2in.

[9] 甲野佑, 高橋達二 (2010), “緩い対称性モデルにおける不確定情報の扱い”, 日本認知科学第*28*回大会, P2-12.

[10] R. S. Sutton(1996), “Generalization in Reinforcement Learning: Successful Examples Using Sparse Coarse Coding”, *Advances in Neural Information Processing Systems 8, MIT Press*, 1038–1044.

[11] 清水 隆宏, 横川 純貴, 甲野 佑, 高橋 達二 (2011), “認知バイアス調整機構 LS の Q 学習への実装とその機能”, 2011年度人工知能学会全国大会, *2011*年度人工知能学会全国大会*(*第*25*回*)*予稿集, 1P2-12in.

[12] D. Uragami, T. Takahashi, H.Alsubeheen, A. Sekiguchi and Y. Matsuo (2011), “The Efficacy of Symmetric Cognitive Biases in Robotic Motion Learning”. *Proceedings of the IEEE ICMA2011* August 7–10, Beijing, China, 410–415.